

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
1 August 2002 (01.08.2002)

PCT

(10) International Publication Number
WO 02/059354 A2

(51) International Patent Classification⁷: **C12Q 1/68**

Daniel [CA/CA]; 7 Chaplin Court, Bolton, Ontario L7E 5Y1 (CA).

(21) International Application Number: **PCT/CA02/00087**

(74) Agents: **HUNT, John, C. et al.; Blake, Cassels & Graydon LLP, Box 25, Commerce Court West, Toronto, Ontario M5L 1A9 (CA).**

(22) International Filing Date: 25 January 2002 (25.01.2002)

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(25) Filing Language: English

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

(26) Publication Language: English

Published:

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(30) Priority Data:
60/263,710 25 January 2001 (25.01.2001) US
60/303,799 10 July 2001 (10.07.2001) US

(71) Applicant (*for all designated States except US*): **TM BIOSCIENCE CORPORATION** [CA/CA]; 439 University Avenue, Suite 1100, Toronto, Ontario M5G 1Y8 (CA).

(72) Inventors; and

(75) Inventors/Applicants (*for US only*): **PANCOSKA, Petr** [CZ/US]; 901 Hinman Avenue #2C, Evanston, IL 60202 (US). **JANOTA, Vit** [CZ/CZ]; Ovenceka 27, 170 00 Praha 7 (CZ). **BENIGHT, Albert, S.** [US/US]; 1630 Valley View Drive, Schaumburg, IL 60193 (US). **BULLOCK, Richard, S.** [US/US]; 3500 North Lake Shore Drive, Chicago, IL 60657 (US). **RICCELLI, Peter, V.** [US/US]; 16830 Richards Drive, Tinley Park, IL 60477 (US). **KOBLER, Daniel** [CH/CA]; 33 Wood Street, Apartment 1102, Toronto, Ontario M4Y 2P8 (CA). **FIELDHOUSE,**



WO 02/059354 A2

(54) Title: POLYNUCLEOTIDES FOR USE AS TAGS AND TAG COMPLEMENTS, MANUFACTURE AND USE THEREOF

(57) Abstract: A family of minimally cross-hybridizing nucleotide sequences, methods of use, etc. A specific family of 210 24mers is described.

**POLYNUCLEOTIDES FOR USE AS TAGS AND TAG COMPLEMENTS,
MANUFACTURE AND USE THEREOF**

FIELD OF THE INVENTION

5 This invention relates to families of oligonucleotide tags for use, for example, in sorting molecules. Members of a given family of tags can be distinguished one from the other by specific hybridization to their tag complements.

10 BACKGROUND OF THE INVENTION

Specific hybridization of oligonucleotides and their analogs is a fundamental process that is employed in a wide variety of research, medical, and industrial applications, including the identification of disease-related polynucleotides in diagnostic assays, screening for clones of novel target 15 polynucleotides, identification of specific polynucleotides in blots of mixtures of polynucleotides, therapeutic blocking of inappropriately expressed genes and DNA sequencing. Sequence specific hybridization is critical in the development of high throughput multiplexed nucleic acid assays. As formats for these assays expand to encompass larger amounts of 20 sequence information acquired through projects such as the Human Genome project, the challenge of sequence specific hybridization with high fidelity is becoming increasingly difficult to achieve.

In large part, the success of hybridization using oligonucleotides depends on minimizing the number of false positives and false negatives. 25 Such problems have made the simultaneous use of multiple hybridization probes in a single experiment i.e. multiplexing, particularly in the analysis of multiple gene sequences on a gene microarray, very difficult. For example, in certain binding assays, a number of nucleic acid molecules are bound to a chip with the desire that a given "target" sequence will bind selectively to 30 its complement attached to the chip. Approaches have been developed that involve the use of oligonucleotide tags attached to a solid support that can be used to specifically hybridize to the tag complements that are coupled to probe sequences. Chetverin et al. (WO 93/17126) uses sectioned, binary oligonucleotide arrays to sort and survey nucleic acids. These arrays have a 35 constant nucleotide sequence attached to an adjacent variable nucleotide sequence, both bound to a solid support by a covalent linking moiety. These binary arrays have advantages compared with ordinary arrays in that they can be used to sort strands according to their terminal sequences so that each strand binds to a fixed location on an array. The design of the terminal 40 sequences in this approach comprises the use of constant and variable

- 2 -

sequences. United States Patent Nos. 6,103,463 and 6,322,971 issued to Chetverin et al. on August 15, 2000 and November 27, 2001, respectively.

This concept of using molecular tags to sort a mixture of molecules is analogous to molecular tags developed for bacterial and yeast genetics 5 (Hensel et al., Science; 269, 400-403: 1995 and Schoemaker et al., Nature Genetics; 14, 450-456: 1996). Here, a method termed "signature tagged" mutagenesis in which each mutant is tagged with a different DNA sequence is used to recover mutant genes from a complex mixture of approximately 10,000 bacterial colonies. In the tagging approach of Barany et al. (WO 9731256), 10 known as the "zip chip", a family of nucleic acid molecules, the "zip-code addresses", each different from each other, are set out on a grid. Target molecules are attached to oligonucleotide sequences complementary to the "zipcode addresses," referred to as "zipcodes," which are used to specifically hybridize to the address locations on the grid. While the 15 selection of these families of polynucleotide sequences used as addresses is critical for correct performance of the assay, the performance has not been described.

Working in a highly parallel hybridization environment requiring specific hybridization imposes very rigorous selection criteria for the 20 design of families of oligonucleotides that are to be used. The success of these approaches is dependent on the specific hybridization of a probe and its complement. Problems arise as the family of nucleic acid molecules cross-hybridize or hybridize incorrectly to the target sequences. While it is common to obtain incorrect hybridization resulting in false positives or 25 an inability to form hybrids resulting in false negatives, the frequency of such results must be minimized. In order to achieve this goal certain thermodynamic properties of forming nucleic acid hybrids must be considered. The temperature at which oligonucleotides form duplexes with their 30 complementary sequences known as the T_m (the temperature at which 50% of the nucleic acid duplex is dissociated) varies according to a number of sequence dependent properties including the hydrogen bonding energies of the canonical pairs A-T and G-C (reflected in GC or base composition), stacking free energy and, to a lesser extent, nearest neighbour interactions. These energies vary widely among oligonucleotides that are typically used in hybridization 35 assays. For example, hybridization of two probe sequences composed of 24 nucleotides, one with a 40% GC content and the other with a 60% GC content, with its complementary target under standard conditions theoretically may have a 10°C difference in melting temperature (Mueller et al., Current Protocols in Mol. Biol.; 15, 5:1993). Problems in hybridization occur when 40 the hybrids are allowed to form under hybridization conditions that include a

- 3 -

single hybridization temperature that is not optimal for correct hybridization of all oligonucleotide sequences of a set. Mismatch hybridization of non-complementary probes can occur forming duplexes with measurable mismatch stability (Santalucia et al., Biochemistry; 38: 3468-77, 5 1999). Mismatching of duplexes in a particular set of oligonucleotides can occur under hybridization conditions where the mismatch results in a decrease in duplex stability that results in a higher T_m than the least stable correct duplex of that particular set. For example, if hybridization is carried out under conditions that favor the AT-rich perfect match duplex 10 sequence, the possibility exists for hybridizing a GC-rich duplex sequence that contains a mismatched base having a melting temperature that is still above the correctly formed AT-rich duplex. Therefore design of families of oligonucleotide sequences that can be used in multiplexed hybridization reactions must include consideration for the thermodynamic properties of 15 oligonucleotides and duplex formation that will reduce or eliminate cross hybridization behavior within the designed oligonucleotide set.

A multiplex sequencing method has been described in United States Patent No. 4,942,124, which issued to Church on July 17, 1990. The method requires at least two vectors which differ from each other at a 20 tag sequence. It is stated in the specification that a tag sequence in one vector will not hybridize under stringent hybridization conditions to a tag sequence in another vector, i.e. a complementary probe of a tag in one vector does not cross-hybridize with a tag sequence in another vector. Exemplary stringent hybridization conditions are given 25 as 42°C in 500-1000 mM sodium phosphate buffer. A set of 42 20-mer tag sequences, all of which lack G residues, is given in Figure 3 of Church's specification. Details of how the sequences were obtained are not provided, although Church states that initially 92 were chosen on the basis of their having sufficient sequence diversity to insure 30 uniqueness.

There have been other attempts at the development of families of tags. There are a number of different approaches for selecting sequences for use in multiplexed hybridization assays. The selection of sequences that can be used as zipcodes or tags in an addressable array has been described in the 35 patent literature in an approach taken by Brenner and co-workers. United States Patent No. 5,654,413 describes a population of oligonucleotide tags (and corresponding tag complements) in which each oligonucleotide tag includes a plurality of subunits, each subunit consisting of an oligonucleotide having a length of from three to six nucleotides and each 40 subunit being selected from a minimally cross hybridizing set, wherein a

- 4 -

subunit of the set would have at least two mismatches with any other sequence of the set. Table II of the Brenner patent specification describes exemplary groups of 4mer subunits that are minimally cross hybridizing according to the aforementioned criteria. In the approach taken by Brenner, constructing non 5 cross-hybridizing oligonucleotides, relies on the use of subunits that form a duplex having at least two mismatches with the complement of any other subunit of the same set. The ordering of subunits in the construction of oligonucleotide tags is not specifically defined.

Parameters used in the design of tags based on subunits are discussed 10 in Barany et al. (WO 9731256). For example, in the design of polynucleotide sequences that are for example 24 nucleotides in length (24mer) derived from a set of four possible tetramers in which each 24mer "address" differs from its nearest 24mer neighbour by 3 tetramers. They discuss further that, if each tetramer differs from each other by at least two nucleotides, then each 15 24mer will differ from the next by at least six nucleotides. This is determined without consideration for insertions or deletions when forming the alignment between any two sequences of the set. In this way a unique "zip code" sequence is generated. The zip code is ligated to a label in a target dependent manner, resulting in a unique "zip code" which is then allowed to 20 hybridize to its address on the chip. To minimize cross-hybridization of a "zip code" to other "addresses", the hybridization reaction is carried out at temperatures of 75-80°C. Due to the high temperature conditions for hybridization, 24mers that have partial homology hybridize to a lesser extent than sequences with perfect complementarity and represent 'dead zones'. This 25 approach of implementing stringent hybridization conditions for example, involving high temperature hybridization, is also practiced by Brenner et. al.

The current state of technology for designing non-cross hybridizing tags based on subunits does not provide sufficient guidance to construct a 30 family of sequences with practical value in assays that require stringent non-cross hybridizing behavior.

Thus, while it is desirable with such arrays to have, at once, a large number of address molecules, the address molecules should each be highly selective for its own complement sequence. While such an array provides the 35 advantage that the family of molecules making up the grid is entirely of design, and does not rely on sequences as they occur in nature, the provision of a family of molecules, which is sufficiently large and where each individual member is sufficiently selective for its complement over all the other zipcode molecules (i.e., where there is sufficiently low cross-hybridization, or cross-talk) continues to elude researchers.

SUMMARY OF INVENTION

Using the method of Benight et al. (described in commonly-owned international patent application No. PCT/CA 01/00141 published under 5 WO 01/59151 on August 16, 2001) a family of 100 nucleotide sequences was obtained using a computer algorithm to have optimal hybridization properties for use in nucleic acid detection assays. The sequence set of 100 oligonucleotides was characterized in hybridization assays, demonstrating the ability of family members to correctly hybridize to their complementary 10 sequences with an absence of cross hybridization. These are the sequences having SEQ ID NOS:1 to 100 of Table I. This set of sequences has been expanded to include an additional 110 sequences that can be grouped with the original 100 sequences as having non-cross hybridizing properties, based on the characteristics of the original set of 100 sequences. These additional 15 sequences are identified as SEQ ID NOS:101 to 210 of the sequences in Table I. How these sequences were obtained is described below.

Variant families of sequences (seen as tags or tag complements) of a family of sequences taken from Table I are also part of the invention. For the purposes of discussion, families of tag complements 20 will be described.

A family of complements is obtained from a set of oligonucleotides based on a family of oligonucleotides such as those of Table I. For illustrative purposes, providing a family of complements based on the oligonucleotides of Table I will be described.

25 Firstly, sequences based on the oligonucleotides of Table I can be represented as follows:

Table IA: Numeric sequences corresponding to word patterns of a set of oligonucleotides

Sequence Identifier	Numeric Pattern					
1	1	4	6	6	1	3
2	2	4	5	5	2	3
3	1	8	1	2	3	4
4	1	7	1	9	8	4
5	1	1	9	2	6	9
6	1	2	4	3	9	6
7	9	8	9	8	10	9
8	9	1	2	3	8	10
9	8	8	7	4	3	1
10	1	1	1	1	1	2
11	2	1	3	3	2	2
12	3	1	2	2	3	2
13	4	1	4	4	4	2
14	1	2	3	3	1	1

- 6 -

Table IA: Numeric sequences corresponding to word patterns of a set of oligonucleotides

Sequence Identifier	Numeric Pattern					
15	1	3	2	2	1	4
16	3	3	3	3	3	4
17	4	3	1	1	4	4
18	3	4	1	1	3	3
19	3	6	6	6	3	5
20	6	6	1	1	6	5
21	7	6	7	7	7	5
22	8	7	5	5	8	8
23	2	1	7	7	1	1
24	2	3	2	3	1	3
25	2	6	5	6	1	6
26	4	8	1	1	3	8
27	5	3	1	1	6	3
28	5	6	8	8	6	6
29	8	3	6	5	7	3
30	1	2	3	1	4	6
31	1	5	7	5	4	3
32	2	1	6	7	3	6
33	2	6	1	3	3	1
34	2	7	6	8	3	1
35	3	4	3	1	2	5
36	3	5	6	1	2	7
37	3	6	1	7	2	7
38	4	6	3	5	1	7
39	5	4	6	3	8	6
40	6	8	2	3	7	1
41	7	1	7	8	6	3
42	7	3	4	1	6	8
43	4	7	7	1	2	4
44	3	6	5	2	6	3
45	1	4	1	4	6	1
46	3	3	1	4	8	1
47	8	3	3	5	3	8
48	1	3	6	6	3	7
49	7	3	8	6	4	7
50	3	1	3	7	8	6
51	10	9	5	5	10	10
52	7	10	10	10	7	9
53	9	9	7	7	10	9
54	9	3	10	3	10	3
55	9	6	3	4	10	6
56	10	4	10	3	9	4
57	3	9	3	10	4	9
58	9	10	5	9	4	8
59	3	9	4	9	10	7
60	3	5	9	4	10	8
61	4	10	5	4	9	3
62	5	3	3	9	8	10
63	6	8	6	9	7	10
64	4	6	10	9	6	4
65	4	9	8	10	8	3
66	7	7	9	10	5	3
67	8	8	9	3	9	10
68	8	10	2	9	5	9

- 7 -

Table IA: Numeric sequences corresponding to word patterns of a set of oligonucleotides

Sequence Identifier	Numeric Pattern					
69	9	6	2	2	7	10
70	9	7	5	3	10	6
71	10	3	6	8	9	2
72	10	9	3	2	7	3
73	8	9	10	3	6	2
74	3	2	5	10	8	9
75	8	2	3	10	2	9
76	6	3	9	8	2	10
77	3	7	3	9	9	10
78	9	10	1	1	9	4
79	10	1	9	1	4	1
80	7	1	10	9	8	1
81	9	1	10	1	10	6
82	9	6	9	1	3	10
83	3	10	8	8	9	1
84	3	8	1	9	10	3
85	9	10	1	3	6	9
86	1	9	1	10	3	1
87	1	4	9	6	8	10
88	3	3	9	6	1	10
89	5	3	1	6	9	10
90	6	1	8	10	9	6
91	5	9	9	4	10	3
92	2	10	9	1	9	5
93	10	10	7	2	1	9
94	10	9	9	1	8	2
95	1	8	6	8	9	10
96	1	9	1	3	8	10
97	9	6	9	10	1	2
98	1	10	8	9	9	2
99	1	9	6	7	2	9
100	4	3	9	3	5	1
101	5	11	10	14	12	1
102	7	12	4	13	3	2
103	5	5	4	4	12	9
104	2	13	13	11	13	13
105	10	2	5	4	12	7
106	11	7	4	11	6	4
107	12	12	1	9	11	11
108	12	9	4	14	12	6
109	12	7	13	2	9	11
110	9	11	3	4	1	3
111	10	5	12	11	4	4
112	4	13	7	12	1	5
113	9	13	10	11	11	6
114	10	14	14	10	1	3
115	2	14	1	10	4	5
116	10	12	12	7	11	10
117	9	11	2	12	8	11
118	2	8	5	2	12	14
119	1	8	13	3	7	8
120	9	4	7	5	4	2
121	13	2	12	7	1	12
122	11	10	9	7	5	11

- 8 -

Table IA: Numeric sequences corresponding to
word patterns of a set of
oligonucleotides

Sequence Identifier	Numeric Pattern					
123	8	12	2	2	12	7
124	5	2	14	3	4	13
125	1	8	8	1	5	9
126	14	5	11	10	13	3
127	14	1	4	13	2	4
128	4	4	5	11	3	10
129	10	9	2	3	3	11
130	11	4	8	14	3	4
131	5	1	14	8	11	2
132	14	3	11	6	12	5
133	13	4	4	1	10	1
134	6	10	11	6	5	1
135	5	8	12	5	1	7
136	4	5	9	6	9	2
137	13	2	4	4	2	3
138	11	2	2	5	9	3
139	8	1	10	12	2	8
140	12	7	9	11	4	1
141	12	1	4	14	3	13
142	11	2	7	10	4	1
143	3	4	12	11	11	11
144	3	3	4	2	12	11
145	1	5	9	4	2	1
146	6	1	12	2	10	5
147	10	5	1	12	2	14
148	2	11	7	9	4	11
149	7	4	4	5	14	12
150	12	5	2	1	10	12
151	5	9	2	11	6	1
152	12	14	3	6	1	14
153	5	9	11	10	1	4
154	2	5	12	14	10	10
155	4	5	8	4	5	6
156	10	12	4	6	12	5
157	4	2	1	13	6	8
158	9	10	10	14	5	3
159	6	14	10	11	3	3
160	2	9	10	12	5	7
161	13	3	7	10	5	12
162	6	4	1	2	5	13
163	6	1	13	4	14	13
164	2	12	1	14	1	9
165	4	11	13	2	6	10
166	1	10	7	4	5	8
167	7	2	2	10	13	4
168	8	2	11	4	6	14
169	4	8	2	6	2	3
170	7	1	12	11	2	9
171	5	6	10	4	13	4
172	5	10	4	11	9	3
173	3	11	9	3	2	3
174	8	15	6	20	17	19
175	21	10	15	3	7	11
176	11	7	17	20	14	9

- 9 -

Table IA: Numeric sequences corresponding to word patterns of a set of oligonucleotides

Sequence Identifier	Numeric Pattern					
177	16	6	17	13	21	21
178	10	15	22	6	17	21
179	15	7	17	10	22	22
180	3	20	8	15	20	16
181	17	21	10	16	6	22
182	6	21	14	14	14	16
183	7	17	3	20	10	7
184	16	19	14	17	7	21
185	20	16	7	15	22	10
186	20	10	18	11	22	18
187	18	7	19	15	7	22
188	21	18	7	21	16	3
189	14	13	7	22	17	13
190	19	7	8	12	10	17
191	15	3	21	14	9	7
192	19	6	15	7	14	14
193	4	17	10	15	20	19
194	21	6	18	4	20	16
195	2	19	8	17	6	13
196	12	12	6	17	4	20
197	16	21	12	10	19	16
198	14	14	15	2	7	21
199	8	16	21	6	22	16
200	14	17	22	14	17	20
201	10	21	7	15	21	18
202	16	13	20	18	21	12
203	15	7	4	22	14	13
204	7	19	14	8	15	4
205	4	5	3	20	7	16
206	22	18	6	18	13	20
207	19	6	16	3	13	3
208	18	6	22	7	20	18
209	10	17	11	21	8	13
210	7	10	17	19	10	14

Here, each of the numerals 1 to 22 (numeric identifiers)

represents a 4mer and the pattern of numerals 1 to 22 of the sequences in the above list corresponds to the pattern of tetrameric

- 5 oligonucleotide segments present in the oligonucleotides of Table I, which oligonucleotides have been found to be non-cross-hybridizing, as described further in the detailed examples. Each 4mer is selected from the group of 4mers consisting of WWWW, WWWX, WWY, WWXW, WWXX, WWXY, WWYW, WWYX, WWYY, WXWW, WXWX, WXWY, WXXX, WXXY, WXYW, WXYX, WXYY, WYWW, WYWX, WYWY, WYXW, WYXX, WYXY, WYYW, WYYX, WYYY, XWWW, XWWX, XWWY, XWXX, XWXY, XWYW, XWYY, XXWW, XXWX, XXWY, XXXW, XXXX, XXXY, XXYW, XXYX, XYWW, XYWX, XYWY, XYXW, XYXX, XYYX, XYYW, YWWX, YWWY, YWXX, YWXY, YWYY, YXXW, YXWX, YXWY,
- 10 XWYW, XYXY, XXYY, XYWW, XYWX, XYWY, XYXW, XYXX, XYYX, XYYW, YWWX, YWWY, YWXX, YWXY, YWYY, YXXW, YXWX, YXWY,

- 10 -

YXXW, YXXX, YXXY, YXYW, YXYX, YXXX, YYWW, YYWX, YYWY, YYXW, YYXX, YYXY,
YYYW, YYYX, and YYYY. Here W, X and Y represent nucleotide bases, A,
G, C, etc., the assignment of bases being made according to rules
described below.

Given this numeric pattern, a 4mer is assigned to a numeral. For
example, 1 = WXYY, 2 = YWXY, etc. Once a given 4mer has been assigned
to a given numeral, it is not assigned for use in the position of a
different numeral. It is possible, however, to assign a different 4mer
to the same numeral. That is, for example, the numeral 1 in one
position could be assigned WXYY and another numeral 1, in a different
position, could be assigned XXXW, but none of the other numerals 2 to
10 can then be assigned WXYY or XXXW. A different way of saying this
is that each of 1 to 22 is assigned a 4mer from the list of eighty-one
4mers indicated so as to be different from all of the others of 1 to
22.

In the case of the specific oligonucleotides given in Table I, 1
= WXYY, 2 = YWXY, 3 = XXXW, 4 = YWYX, 5 = WYXY, 6 = YYWX, 7 = YWXX, 8 =
WYXX, 9 = XYYW, 10 = XYWX, 11 = YYXW, 12 = WYYX, 13 = XYXW, 14 = WYYY,
15 = WXYW, 16 = WYXW, 17 = WXXW, 18 = WYYW, 19 = XYYX, 20 = YXYX, 21 =
YXXX and 22 = XYXY.

Once the 4mers are assigned to positions according to the above
pattern, a particular set of oligonucleotides can be created by
appropriate assignment of bases, A, T/U, G, C to W, X, Y. These
assignments are made according to one of the following two sets of
rules:

(i) Each of W, X and Y is a base in which:

- (a) W = one of A, T/U, G, and C,
X = one of A, T/U, G, and C,
Y = one of A, T/U, G, and C,
and each of W, X and Y is selected so as to be different
from all of the others of W, X and Y, and
- (b) an unselected said base of (i) (a) can be substituted any
number of times for any one of W, X and Y.

or

(ii) Each of W, X and Y is a base in which:

- (a) W = G or C,
X = A or T/U,

- 11 -

Y = A or T/U,

and X ≠ Y, and

- (b) a base not selected in (ii)(a) can be inserted into each sequence at one or more locations, the location of each insertion being the same in each sequence as that of every other ~~sequence~~ of the set.

In the case of the specific oligonucleotides given in Table I, W = G, X = A and Y = T.

In any case, given a set of oligonucleotides generated according to one of these sets of rules, it is possible to modify the members of a given set in relatively minor ways and thereby obtain a different set of sequences while more or less maintaining the cross-hybridization properties of the set subject to such modification. In particular, it is possible to insert up to 3 of A, T/U, G and C at any location of any sequence of the set of sequences. Alternatively, or additionally, up to 3 bases can be deleted from any sequence of the set of sequences.

A person skilled in the art would understand that given a set of oligonucleotides having a set of properties making it suitable for use as a family of tags (or tag complements) one can obtain another family with the same property by reversing the order of all of the members of the set. In other words, all the members can be taken to be read 5' to 3' or to be read 3' to 5'.

A family of complements of the present invention is based on a given set of oligonucleotides defined as described above. Each complement of the family is based on a different oligonucleotide of the set and each complement contains at least 10 consecutive (i.e., contiguous) bases of the oligonucleotide on which it is based. When selecting a sequence of contiguous bases, preference is given to those sets in which the contiguous bases of each oligonucleotide of a set are selected such that the position of the first base of each said oligonucleotide within the sequence on which it is based is the same for all nucleotides of the set. Thus, for example, if a nucleotide sequence of twenty contiguous bases corresponds to bases 3 to 22 of the sequence on which the nucleotide sequence is based, then preferably, the twenty contiguous bases for all nucleotide sequences corresponds to bases 3 to 22 of the sequences on which the nucleotides sequences are based. For a given family of complements where one is seeking to reduce or minimize inter-sequence similarity that would result in cross-hybridization, each and every pair of complements meets

- 12 -

particular homology requirements. Particularly, subject to limited exceptions, described below, any two complements within a set of complements are generally required to have a defined amount of dissimilarity.

5 In order to notionally understand these requirements for dissimilarity as they exist for a given pair of complements of a family, a phantom sequence is generated from the pair of complements. A "phantom" sequence is a single sequence that is generated from a pair 10 of complements by selection, from each complement of the pair, of a string of bases wherein the bases of the string occur in the same order in both complements. An object of creating such a phantom sequence is to create a convenient and objective means of comparing the sequence identity of the two parent sequences from which the phantom sequence is created.

15 A phantom sequence can be considered to be similar in concept to a consensus sequence which a person skilled in the art would be familiar with, except that a consensus sequence typically is comprised of all bases from both parent sequences with each position reflecting the most common choice of base at each position (the union of both 20 sequences), whereas the "phantom" sequence is comprised of only bases which occur in the same order in both parent sequences (the intersection of both sequences). Also, a consensus sequence usually is indicative of a common phylogenetic ancestry for the two sequences (or more than 2 sequences depending on how many sequences are used to 25 generate the consensus sequence), whereas the "phantom" sequence definition has been created to specifically address the sequence similarity between 2 complementary sequences which have no ancestral history but may have a propensity to cross-hybridize under certain conditions.

30 A phantom sequence may thus be generated from exemplary Sequence 1 and Sequence 2 as follows:

Sequence 1: ATGTTTAGTGAAAGTTAGTATTG

*

•

Sequence 2: ATGTTAGTGAATAGTATAGTATTG

•

◆

Phantom Sequence: ATGTTAGTGAAGTTAGTATTG

- 13 -

The phantom sequence generated from these two sequences is thus 22 bases in length. That is, one can see that there are 22 identical bases with identical sequence (the same order) in Sequence Nos. 1 and 2. There is a total of three insertions/deletions and mismatches 5 present in the phantom sequence when compared with the sequences from which it was generated:

ATGT-TAGTGAA-AGT-TAGTATTG

10 The dashed lines in this latter representation of the phantom sequence indicate the locations of the insertions/deletions and mismatches in the phantom sequence relative to the parent sequences from which it was derived. Thus, the "T" marked with an asterisk in Sequence 1, the "A" marked with a diamond in Sequence 2 and the "A-T" mismatch of Sequences 15 1 and 2 marked with two dots were deleted in generating the phantom sequence.

A person skilled in the art will appreciate that the term "insertion/deletion" is intended to cover the situations indicated by the asterisk and diamond. Whether the change is considered, strictly 20 speaking, an insertion or deletion is merely one of vantage point. That is, one can see that the fourth base of Sequence 1 can be deleted therefrom to obtain the phantom sequence, or a "T" can be inserted after the third base of the phantom sequence to obtain Sequence 1.

One can thus see that if it were possible to create a phantom 25 sequence by elimination of a single insertion/deletion from one of the parent sequences, that the two parent sequences would have identical homology over the length of the phantom sequence except for the presence of a single base in one of the two sequences being compared. Likewise, one can see that if it were possible to create a phantom 30 sequence through deletion of a mismatched pair of bases, one base in each parent, that the two parent sequences would have identical homology over the length of the phantom sequence except for the presence of a single base in each of the sequences being compared. For this reason, the effect of an insertion/deletion is considered 35 equivalent to the effect of a mismatched pair of bases when comparing the homology of two sequences.

Once a phantom sequence is generated, the compatibility of the pair of complements from which it was generated within a family of complements can be systematically evaluated.

- 14 -

According to one embodiment of the invention, a pair of complements is compatible for inclusion within a family of complements if any phantom sequence generated from the pair of complements has the following properties:

5

- (1) Any consecutive sequence of bases in the phantom sequence which is identical to a consecutive sequence of bases in each of the first and second complements from which it is generated is no more than $((3/4 \times L) - 1)$ bases in length;
- (2) The phantom sequence, if greater than or equal to $(5/6 \times L)$ in length, contains at least 3 insertions/deletions or mismatches when compared to the first and second complements from which it is generated; and
- (3) The phantom sequence is not greater than or equal to $(11/12 \times L)$ in length.

Here, L_1 is the length of the first complement, L_2 is the length of the second complement, and $L = L_1$, or if $L_1 \neq L_2$, L is the greater of L_1 and L_2 .

10

In particular preferred embodiments of the invention, all pairs of complements of a given set have the properties set out above. Under particular circumstances, it may be advantageous to have a limited number of complements that do not meet all of these requirements when compared to every other complement in a family.

15

In one case, for any first complement there are at most two second complements in the family which do not meet all of the three listed requirements. For two such complements, there would thus be a greater chance of cross-hybridization between their tag counterparts and the first complement. In another case, for any first complement there is at most one second complement which does not meet all of three listed requirements.

20

It is also possible, given this invention, to design a family of complements where a specific number or specific portion of the complements do not meet the three listed requirements. For example, a set could be designed where only one pair of complements within the set do not meet the requirements when compared to each other. There could be two pairs, three pairs, and any number of pairs up to and including all possible pairs. Alternatively, it may be advantageous to have a given proportion of pairs of complements that do not meet the

- 15 -

requirements, say 10% of pairs, when compared with other sequences that do not meet one or more of the three requirements listed. This number could instead be 5%, 15%, 20%, 25%, 30%, 35%, or 40%.

5 The foregoing comparisons would generally be largely carried out using appropriate computer software. Although notionally described in terms of a phantom sequence for the sake of clarity and understanding, it will be understood that a competent computer programmer can carry out pairwise comparisons of complements in any number of ways using logical steps that obtain equivalent results.

10 The symbols A, G, T/U, C take on their usual meaning in the art here. In the case of T and U, a person skilled in the art would understand that these are equivalent to each other with respect to the inter-strand hydrogen-bond (Watson-Crick) binding properties at work in the context of this invention. The two bases are thus interchangeable
15 and hence the designation of T/U.

Analogues of the naturally occurring bases can be inserted in
their respective places where desired. An Analogue is any non-natural
base, such as peptide nucleic acids and the like that undergoes normal
Watson-Crick pairing in the same way as the naturally occurring
20 nucleotide base to which it corresponds.

In one broad aspect, the present invention is thus a composition comprising molecules for use as tags or tag complements wherein each molecule comprises an oligonucleotide selected from a set of oligonucleotides based on a group of sequences having numeric patterns
25 as set out in Table IA wherein:

- (A) each of 1 to 22 is a 4mer selected from the group of 4mers consisting of WWWW, WWWX, WWWY, WWXW, WWXX, WWXY, WWYW, WWYX, WWYY, WXWW, WXWX,
WXWY, WXXX, WXXX, WXXY, WXYW, WXYX, WXYY, WYWW, WYWX, WYWY, WYXW,
WYXX, WYXY, WYYW, WYYX, WYYY, XWWW, XWWX, XWWY, XWXW, XWXX, XWXY,
XWYW, XWYY, XXWW, XXWX, XXWY, XXXW, XXXX, XXXY, XXYW, XXYX,
XXYY, XYWW, XYWX, XYWY, XYXW, XYXX, XYXY, XYYW, XYYX, XYYY, YWWW,
YWWX, YWWY, YWXW, YWXX, YWXY, YWYW, YWYX, YWYY, YXWW, YXWX, YXWY,
YXXW, YXXX, YXXX, YXYW, YXYX, YXYY, YYWW, YYWX, YYWY, YYXW, YYXX,
YYXY, YYYW, YYYX, and YYYY, and
- (B) each of 1 to 22 is selected so as to be different from all of the others of 1 to 22;
- (C) each of W, X and Y is a base in which either (i) or (ii) is true:
 - (i) (a) W = one of A, T/U, G, and C,
X = one of A, T/U, G, and C,
Y = one of A, T/U, G, and C,

- 16 -

and each of W, X and Y is selected so as to be different from all of the others of W, X and Y, and

- (b) an unselected said base of (i)(a) can be substituted any number of times for any one of W, X and Y,

- (ii) (a) W = G or C,

X = A or T/U,

Y = A or T/U,

and X ≠ Y, and

- (b) a base not selected in (ii)(a) can be inserted into each sequence at one or more locations, the location of each insertion being the same in all the sequences;

- (D) up to three bases can be inserted at any location of any of the sequences or up to three bases can be deleted from any of the sequences;

- (E) all of the sequences of a said group of oligonucleotides are read 5' to 3' or are read 3' to 5'; and

wherein each oligonucleotide of a said set has a sequence of at least ten contiguous bases of the sequence on which it is based, provided that:

- (F) (I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.1 and 0.40 and said quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.2; and

- (II) for any phantom sequence generated from any pair of first and second sequences of the set L₁ and L₂ in length, respectively, by selection from the first and second sequences of identical bases in identical sequence with each other:

- (i) any consecutive sequence of bases in the phantom sequence which is identical to a consecutive sequence of bases in each of the first and second sequence from which it is generated is less than ((3/4 × L) - 1) bases in length;

- (ii) the phantom sequence, if greater than or equal to (5/6 × L) in length, contains at least three insertions/deletions or mismatches when compared to the first and second sequences from which it is generated; and

- (iii) the phantom sequence is not greater than or equal to (11/12 × L) in length;

- 17 -

where $L = L_1$, or if $L_1 \neq L_2$, where L is the greater of L_1 and L_2 ;
and

wherein any base present may be substituted by an analogue thereof.

In a preferred embodiment, a set of oligonucleotides of the invention is based on the numeric patterns of sequences tested in Example 2.

Preferably,

(G) for the group of 24mer sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, for the group of sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, under a defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 30% of the degree of hybridization between said sequence and its complement, for all oligonucleotides of the set, the maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 50% of the degree of hybridization of the oligonucleotide and its complement.

It can thus be seen that it is possible to routinely determine

- 5 whether all oligonucleotides of a selected set are all minimally cross-hybridizing. Preferably in (G), under said defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence does not exceed 30% of the degree of hybridization between said sequence and its complement, it is also true
10 that the degree of hybridization between each sequence and its complement varies by a factor of between 1 and 10, more preferably between 1 and 9, and more preferably between 1 and 8. It is demonstrated in Example 2, below, for a preferred set of oligonucleotides, that the degree of hybridization between each
15 sequence and its specific complement varies by a factor of between 1 and 8.25 and the maximum degree of hybridization between a sequence and any complement of a different sequence does not exceed 10.2% of the

- 18 -

degree of hybridization between the sequence and its specific complement.

Preferably, the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 25%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 20%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 15%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 11%.

Preferably, under the defined set of conditions of (G), the maximum degree of hybridization between a sequence and a complement of any other sequence of the set is no more than 15% greater than the maximum degree of hybridization between a sequence and any complement of a different sequence of the said group of 24mer sequences, more preferably no more than 10% greater, more preferably no more than 5% greater.

According to Example 2, described below, under conditions of 0.2 M NaCl, 0.1 M Tris, 0.08% Triton X-100, pH 8.0 at 37°C, the maximum degree of hybridization between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 10.2% when 24mer nucleotide sequences are covalently linked to a solid support, in this case microparticles or beads.

In another preferred aspect of the composition, in (G) for the group of 24mers the maximum degree of hybridization between a sequence and any complement of a different sequence does not exceed 15% of the degree of hybridization between said sequence and its complement and the degree of hybridization between each sequence and its complement varies by a factor of between 1 and 9, and for all oligonucleotides of the set, the maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 20% of the degree of hybridization of the oligonucleotide and its complement.

In a preferred aspect, each of the 4mers represented by numerals 1 to 22 is selected from the group of 4mers consisting of WXXX, WXXY, WXYX, WYYY, WYXX, WYXY, WYYX, WYYY, XWXX, XWXY, XWYY, XXWX, XXWY,

- 19 -

XXXW, XXYW, XYWX, XYWY, XYXW, XYYW, YWXX, YWXY, YWYY, YXWX, YXWY, YXXW, YXYW, YYWX, YYWY, YYXW, and YYYW.

In another aspect, each of the 4mers represented by numeral 1 are identical to each other, each of the 4mers represented by numeral 2 are 5 identical to each other, each of the 4mers represented by numeral 3 are identical to each other, each of the 4mers represented by numeral 4 are identical to each other, each of the 4mers represented by numeral 5 are identical to each other, each of the 4mers represented by numeral 6 are identical to each other, each of the 4mers represented by numeral 7 are 10 identical to each other, each of the 4mers represented by numeral 8 are identical to each other, each of the 4mers represented by numeral 9 are identical to each other, each of the 4mers represented by numeral 10 are identical to each other, each of the 4mers represented by numeral 11 are identical to each other, each of the 4mers represented by 15 numeral 12 are identical to each other, each of the 4mers represented by numeral 13 are identical to each other, each of the 4mers represented by numeral 14 are identical to each other, each of the 4mers represented by numeral 15 are identical to each other, each of the 4mers represented by numeral 16 are identical to each other, each 20 of the 4mers represented by numeral 17 are identical to each other, each of the 4mers represented by numeral 18 are identical to each other, each of the 4mers represented by numeral 19 are identical to each other, each of the 4mers represented by numeral 20 are identical to each other, each of the 4mers represented by numeral 21 are 25 identical to each other, and each of the 4mers represented by numeral 22 are identical to each other.

In another aspect, at least one of the 4mers represented by the numeral 1 has the sequence WXYY, at least one of the 4mers represented by the numeral 2 has the sequence YWXY, at least one of the 4mers 30 represented by the numeral 3 has the sequence XXXW, at least one of the 4mers represented by the numeral 4 has the sequence YWYX, at least one of the 4mers represented by the numeral 5 has the sequence WYXY, at least one of the 4mers represented by the numeral 6 has the sequence YYWX, at least one of the 4mers represented by the numeral 7 has the 35 sequence YWXX, at least one of the 4mers represented by the numeral 8 has the sequence WYXX, at least one of the 4mers represented by the numeral 9 has the sequence XYYW, at least one of the 4mers represented by the numeral 10 has the sequence XYWX, at least one of the 4mers represented by the numeral 11 has the sequence YYXW, at least one of the 40 4mers represented by the numeral 12 has the sequence WYYX, at least

- 20 -

one of the 4mers represented by the numeral 13 has the sequence XYXW,
at least one of the 4mers represented by the numeral 14 has the
sequence WYYY, at least one of the 4mers represented by the numeral 15
has the sequence WXYW, at least one of the 4mers represented by the
5 numeral 16 has the sequence WYXW, at least one of the 4mers represented
by the numeral 17 has the sequence WXXW, at least one of the 4mers
represented by the numeral 18 has the sequence WYYW, at least one of
the 4mers represented by the numeral 19 has the sequence XYYX, at least
one of the 4mers represented by the numeral 20 has the sequence YXYX,
10 at least one of the 4mers represented by the numeral 21 has the
sequence YXXX, and/or at least one of the 4mers represented by the
numeral 22 has the sequence XYXY.

In one preferred aspect, the invention is a composition in which
each 1 = WXYY, each 2 = YWXY, each 3 = XXXW, each 4 = YWYX, each 5 =
15 WYXY, each 6 = YYWX, each 7 = YWXX, each 8 = WYXX, each 9 = XYYW, each
10 = XYWX, each 11 = YYXW, each 12 = WYYX, each 13 = XYXW, each 14 =
WYYY, each 15 = WXYW, each 16 = WYXW, each 17 = WXXW, each 18 = WYYW,
each 19 = XYYX, each 20 = YXYX, each 21 = YXXX and each 22 = XYXY.

In one broad aspect, the invention is a composition wherein a
20 group of sequences is based on those having numeric patterns of those
with numeric identifiers 1 to 173 of Table IA and wherein each of the
4mers represented by numerals 1 to 14 in (A) is selected from the group
of 4mers consisting of WXYY, YWXY, XXXW, YWYX, WYXY, YYWX, YWXX, WYXX,
XYYW, XYWX, YYXW, WYYX, XYXW, and WYYY.

25 In such a composition it is preferred that each of the 4mers
represented by numeral 1 are identical to each other, each of the 4mers
represented by numeral 2 are identical to each other, each of the 4mers
represented by numeral 3 are identical to each other, each of the 4mers
represented by numeral 4 are identical to each other, each of the 4mers
represented by numeral 5 are identical to each other, each of the 4mers
represented by numeral 6 are identical to each other, each of the 4mers
represented by numeral 7 are identical to each other, each of the 4mers
represented by numeral 8 are identical to each other, each of the 4mers
represented by numeral 9 are identical to each other, each of the 4mers
represented by numeral 10 are identical to each other, each of the
30 4mers represented by numeral 11 are identical to each other, each of
the 4mers represented by numeral 12 are identical to each other, each
of the 4mers represented by numeral 13 are identical to each other,
and/or each of the 4mers represented by numeral 14 are identical to
35 each other.

- 21 -

It is also preferred that at least one of the 4mers represented by the numeral 1 has the sequence WXYY, at least one of the 4mers represented by the numeral 2 has the sequence YWXY, at least one of the 4mers represented by the numeral 3 has the sequence XXXW, at least one 5 of the 4mers represented by the numeral 4 has the sequence YWYX, at least one of the 4mers represented by the numeral 5 has the sequence WYXY, at least one of the 4mers represented by the numeral 6 has the sequence YYWX, at least one of the 4mers represented by the numeral 7 has the sequence YWXX, at least one of the 4mers represented by the numeral 8 has the sequence WYXX, at least one of the 4mers represented by the numeral 9 has the sequence XYYW, at least one of the 4mers 10 represented by the numeral 10 has the sequence XYWX, at least one of the 4mers represented by the numeral 11 has the sequence YYXW, at least one of the 4mers represented by the numeral 12 has the sequence WYYX, 15 at least one of the 4mers represented by the numeral 13 has the sequence XYXW, and/or at least one of the 4mers represented by the numeral 14 has the sequence WYYY.

More preferably, each 1 = WXYY, each 2 = YWXY, each 3 = XXXW, each 4 = YWYX, each 5 = WYXY, each 6 = YYWX, each 7 = YWXX, each 8 = 20 each 9 = XYYW, each 10 = XYWX, each 11 = YYXW, each 12 = WYYX, each 13 = XYXW, and each 14 = WYYY.

In another broad aspect, the invention is a composition in which a group of sequences is based on those sequences having the numeric patterns of those with sequence identifiers 1 to 100 set out in Table IA 25 and wherein each of the 4mers represented by numerals 1 to 10 in (A) is selected from the group of 4mers consisting of WXYY, YWXY, XXXW, YWYX, WYXY, YYWX, YWXX, WYXX, XYYW, and XYWX.

In such a composition it is preferred that each of the 4mers represented by numeral 1 are identical to each other, each of the 4mers 30 represented by numeral 2 are identical to each other, each of the 4mers represented by numeral 3 are identical to each other, each of the 4mers represented by numeral 4 are identical to each other, each of the 4mers represented by numeral 5 are identical to each other, each of the 4mers represented by numeral 6 are identical to each other, each of the 4mers 35 represented by numeral 7 are identical to each other, each of the 4mers represented by numeral 8 are identical to each other, each of the 4mers represented by numeral 9 are identical to each other, and/or each of the 4mers represented by numeral 10 are identical to each other.

It also preferred that at least one of the 4mers represented by 40 the numeral 1 has the sequence WXYY, at least one of the 4mers

- 22 -

represented by the numeral 2 has the sequence YWXY, at least one of the 4mers represented by the numeral 3 has the sequence XXXW, at least one of the 4mers represented by the numeral 4 has the sequence YWYX, at least one of the 4mers represented by the numeral 5 has the sequence WYXY, at least one of the 4mers represented by the numeral 6 has the sequence YYWX, at least one of the 4mers represented by the numeral 7 has the sequence YWXX, at least one of the 4mers represented by the numeral 8 has the sequence WYXX, at least one of the 4mers represented by the numeral 9 has the sequence XYYW, and/or at least one of the 4mers represented by the numeral 10 has the sequence XYWX.

More preferably, each 1 = WXYY, each 2 = YWXY, each 3 = XXXW, each 4 = YWYX, each 5 = WYXY, each 6 = YYWX, each 7 = YWXX, each 8 = WYXX, each 9 = XYYW, and each 10 = XYWX.

In the most preferred compositions, in (C)(i)(a): W = one of G and C; X = one of A and T/U; and Y = one of A and T/U, maintaining the provisos of (F). More preferably, (C)(i)(a): W = G; X = one of A, and T/U; and Y = one of A and T/U. Even more preferably, wherein W = G; X = A; and Y = T/U.

A person skilled in the art will appreciate that the closer a given oligonucleotide sequence variant is to one of the most preferred sequences (Table I), the more closely it will resemble the preferred sequence as a member of a minimally cross-hybridizing set of oligonucleotides.

It will be understood that when it is stated herein that a group of sequences (oligonucleotides) is minimally cross-hybridizing, it is meant that any given member of the group of sequences (oligonucleotides) only minimally hybridizes with the complement of any other sequence (oligonucleotide) of that group.

Preferably, in (F)(I), the quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.1, more preferably, the quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.05, more preferably the quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.01.

Also, it is preferred in (F)(I) that the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.15 and 0.35, more preferably between about 0.2 and 0.3, more preferably between about 0.21 and 0.29, more preferably between about 0.22 and 0.28, more preferably between about

- 23 -

0.23 and 0.27, even more preferably between about 0.24 and 0.26, and most preferably the quotient is 0.25.

Preferably, in (D) up to two bases can be inserted at any location of any of the sequences or up to two bases can be deleted from 5 any of the sequences, more preferably only one base can be inserted at any location of any of the sequences or one base can be deleted from any of the sequences, and most preferably no base is inserted at any location of any of the sequences.

Also, it is preferred that in (D), no base can be deleted from 10 any of the sequences, and most preferably, in (D) no base can be inserted at or deleted from any location of any of the sequences.

In preferred compositions, each of the oligonucleotides of a set has a sequence at least eleven contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a 15 set has a sequence at least twelve contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least thirteen contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least fourteen contiguous bases of the 20 sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least fifteen contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least sixteen contiguous bases of the sequence on which it is based; or more 25 preferably each of the oligonucleotides of a set has a sequence at least seventeen contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least eighteen contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least nineteen contiguous bases of the sequence on which it 30 is based; or more preferably each of the oligonucleotides of a set has a sequence at least twenty contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least twenty-one contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least twenty-two contiguous bases of the sequence on which it is based; or more preferably each of the oligonucleotides of a set has a sequence at least twenty-three contiguous bases of the 35 sequence on which it is based; or more preferably each of the

- 24 -

oligonucleotides of a set has a sequence at least twenty-four contiguous bases of the sequence on which it is based.

Preferably, each of the oligonucleotides of a set is up to thirty bases in length; or more preferably each of the oligonucleotides of a set is up to twenty-nine bases in length; or more preferably each of the oligonucleotides of a set is up to twenty-eight bases in length; or more preferably each of the oligonucleotides of a set is up to twenty-seven bases in length; or more preferably each of the oligonucleotides of a set is up to twenty-six bases in length; or more preferably each of the oligonucleotides of a set is up to twenty-five bases in length; or more preferably each of the oligonucleotides of a set is up to twenty-four bases in length.

In certain preferred embodiments, each of the oligonucleotides of a set has a length of within five bases of the average length of all of the oligonucleotides in the set; or more preferably each of the oligonucleotides of a set has a length of within four bases of the average length of all of the oligonucleotides in the set; or more preferably each of the oligonucleotides of a set has a length of within three bases of the average length of all of the oligonucleotides in the set; or more preferably each of the oligonucleotides of a set has a length of within two bases of the average length of all of the oligonucleotides in the set; or more preferably each of the oligonucleotides of a set has a length of within one base of the average length of all of the oligonucleotides in the set.

Preferably, the string of contiguous bases of each oligonucleotide of a said set are selected such that the position of the first base of each string within the sequence on which it is based is the same for all nucleotides of the set.

In preferred embodiments, the composition includes at least ten said molecules, or at least eleven said molecules, or at least twelve said molecules, or at least thirteen said molecules, or at least fourteen said molecules, or at least fifteen said molecules, or at least sixteen said molecules, or at least seventeen said molecules, or at least eighteen said molecules, or at least nineteen said molecules, or at least twenty said molecules, or at least twenty-one said molecules, or at least twenty-two said molecules, or at least twenty-three said molecules, or at least twenty-four said molecules, or at least twenty-five said molecules, or at least twenty-six said molecules, or at least twenty-seven said molecules, or at least twenty-eight said molecules, or at least twenty-nine said molecules, or at

- 25 -

least thirty said molecules, or at least thirty-one said molecules, or
at least thirty-two said molecules, or at least thirty-three said
molecules, or at least thirty-four said molecules, or at least thirty-
five said molecules, or at least thirty-six said molecules, or at least
5 thirty-seven said molecules, or at least thirty-eight said molecules,
or at least thirty-nine said molecules, or at least forty said
molecules, or at least forty-one said molecules, or at least forty-two
said molecules, or at least forty-three said molecules, or at least
forty-four said molecules, or at least forty-five said molecules, or at
10 least forty-six said molecules, or at least forty-seven said molecules,
or at least forty-eight said molecules, or at least forty-nine said
molecules, or at least fifty said molecules, or at least sixty said
molecules, or at least seventy said molecules, or at least eighty said
molecules, or at least ninety said molecules, or at least one hundred
15 said molecules, or at least, depending upon the size of the group of
sequences on which the oligonucleotides are based, one hundred and ten
said molecules, or at least one hundred and twenty said molecules, or
at least one hundred and thirty said molecules, or at least one hundred
and forty said molecules, or at least one hundred and fifty said
20 molecules, or at least one hundred and sixty said molecules, or at
least one hundred and seventy said molecules, or at least one hundred
and eighty said molecules, or at least one hundred and ninety said
molecules, or at least two hundred said molecules.

A person skilled in the art will appreciate that, depending upon
25 the use to which a family of oligonucleotides of the invention are to
be put, it may or may not be desirable to include with sequences that
can be distinguished one from the other (i.e., are minimally cross-
hybridizing) a number of sequences that do cross hybridize with each
other.

30 In a preferred aspect, the invention is a composition wherein in
(II) (i), any consecutive sequence of bases in the phantom sequence
which is identical to a consecutive sequence of bases in each of the
first and second sequences from which it is generated is no more than
 $((2/3 \times L) - 1)$ bases in length. More preferably, the phantom
35 sequence, if greater than or equal to $(3/4 \times L)$ in length, contains at
least 3 insertions/deletions or mismatches when compared to the first
and second sequences from which it is generated, and even more
preferably, the phantom sequence, if greater than or equal to $(2/3 \times L)$
in length, contains at least 3 insertions/deletions or mismatches when
40 compared to the first and second sequences from which it is generated.

- 26 -

In another preferred aspect, in (II)(iii), the phantom sequence is not greater than or equal to $(5/6 \times L)$ in length, more preferably, the phantom sequence is not greater than or equal to $(3/4 \times L)$ in length.

5 In another broad aspect, the invention is a composition containing molecules for use as tags or tag complements wherein each molecule comprises an oligonucleotide selected from a set of oligonucleotides based on a group of sequences having the numeric patterns of the sequences tested in Example 2, as set out in Table IA,
10 wherein:

(A) wherein 1 = WXYY, each 2 = YWXY, each 3 = XXXW, each 4 = YWYX, each 5 = WYXY, each 6 = YYWX, each 7 = YWXX, each 8 = WYXX, each 9 = XYYW, each 10 = XYWX, each 11 = YYXW, each 12 = WYYX, each 13 = XYXW, each 14 = WYYY, each 15 = WXYW, each 16 = WYXW, each 17 = WXXW, each 18 = WYYW, each 19 = XYYX, each 20 = YXYX, each 21 = YXXX and each 22 = XYXY;

(B) each of W, X and Y is a base in which either:

(i) (a) W = one of A, T/U, G, and C,
X = one of A, T/U, G, and C,
Y = one of A, T/U, G, and C,

and each of W, X and Y is selected so as to be different from all of the others of W, X and Y,

(b) an unselected said base of (i)(a) can be substituted any number of times for any one of W, X and Y, or

(ii) (a) W = G or C,
X = A or T/U,
Y = A or T/U,
and X ≠ Y, and

(b) a base not selected in (ii)(a) can be inserted into each sequence at one or more locations, the location of each insertion being the same in all the sequences;

(C) up to three bases can be inserted at any location of any of the sequences or up to three bases can be deleted from any of the sequences;

(D) all of the sequences of a said group of oligonucleotides are read 5' to 3' or are read 3' to 5'; and

wherein each oligonucleotide of a said set has a sequence of at least ten contiguous bases of the sequence on which it is based, provided that:

(E) the quotient of the sum of G and C divided by the sum of A, T/U, G and

- 27 -

C for all combined sequences of the set is between about 0.1 and 0.40 and said quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.2; and

(F) for the group of 24mer sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, for the group of sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, under a defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 30% of the degree of hybridization between said sequence and its complement, for all oligonucleotides of the set, the maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 50% of the degree of hybridization of the oligonucleotide and its complement; wherein any base present may be substituted by an analogue thereof.

Again, preferably, the contiguous bases of each oligonucleotide of a set are selected such that the position of the first base of each oligonucleotide within the sequence on which it is based is the same for all nucleotides of the set.

5 In a preferred aspect, subject to the provisos of (E) and (F) above, each oligonucleotide of a said set comprises a said sequence of twenty-four contiguous bases of the sequence on which it is based.

More preferably, subject to the proviso of (F) each oligonucleotide of a said set comprises a said sequence of twenty-four contiguous bases of the sequence on which it is based.

In particularly preferred aspects, in (B), W = one of G and C; X = one of A and T/U; and Y = one of A and T/U.

Even more preferred, in (B): W = G; X = one of A, and T/U; and Y = one of A and T/U.

15 In another broad aspect, the invention is a composition that includes fifty minimally cross-hybridizing molecules for use as tags or tag complements wherein each molecule comprises an oligonucleotide

- 28 -

comprising a sequence of nucleotide bases for which, under a defined set of conditions, the maximum degree of hybridization between a said oligonucleotide and any complement of a different oligonucleotide does not exceed about 10% of the degree of hybridization between said 5 oligonucleotide and its complement.

A preferred set of such defined conditions results in a level of hybridization that is the same as the level of hybridization obtained when hybridization conditions include 0.2 M NaCl, 0.1 M Tris, 0.08% Triton X-100, pH 8.0 at 37°C, and the sequences are covalently linked 10 to microparticles. Of course, these conditions are preferably used directly.

Preferably, under the defined set of conditions, whatever the conditions are, the degree of hybridization between each 15 oligonucleotide and its complement varies by a factor of between 1 and 8.

Preferably, each oligonucleotide is the same length and is at least twenty nucleotide bases in length. More preferably, each oligonucleotide is twenty-four nucleotide bases in length.

In certain embodiments, each molecule of a composition is linked 20 to a solid phase support so as to be distinguishable from a mixture of said molecules by hybridization to its complement. Each such molecule can be linked to a defined location on such a solid phase support, the defined location for each molecule being different than the defined location for other, different, molecules.

25 In one preferred embodiment, the solid phase support is a microparticle and each said molecule is covalently attached to a different microparticle than each other different said molecule.

The invention includes kits for sorting and identifying 30 polynucleotides. Such a kit can include one or more solid phase supports each having one or more spatially discrete regions, each such region having a uniform population of substantially identical tag complements covalently attached. The tag complements are made up of a set of oligonucleotides of the invention.

The one or more solid phase supports can be a planar substrate in 35 which the one or more spatially discrete regions is a plurality of spatially addressable regions.

The tag complements can also be coupled to microparticles. Microparticles preferably each have a diameter in the range of from 5 to 40 μm .

- 29 -

Such a kit preferably includes microparticles that are spectrophotometrically unique, and therefore distinguishable from each other according to conventional laboratory techniques. Of course for such kits to work, each type of microparticle would generally have only 5 one tag complement associated with it, and usually there would be a different oligonucleotide tag complement associated with (attached to) each type of microparticle.

The invention includes methods of using families of oligonucleotides of the invention.

10 One such method is of analyzing a biological sample containing a biological sequence for the presence of a mutation or polymorphism at a locus of the nucleic acid. The method includes:

- (A) amplifying the nucleic acid molecule in the presence of a first primer having a 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements of the invention to form an amplified molecule with a 5'-end with a sequence complementary to the sequence of the tag;
- (B) extending the amplified molecule in the presence of a polymerase and a second primer having 5'-end complementary the 3'-end of the amplified sequence, with the 3'-end of the second primer extending to immediately adjacent said locus, in the presence of a plurality of nucleoside triphosphate derivatives each of which is: (i) capable of incorporation during transcription by the polymerase onto the 3'-end of a growing nucleotide strand; (ii) causes termination of polymerization; and (iii) capable of differential detection, one from the other, wherein there is a said derivative complementary to each possible nucleotide present at said locus of the amplified sequence;
- (C) specifically hybridizing the second primer to a tag complement having the tag complement sequence of (A); and
- (D) detecting the nucleotide derivative incorporated into the second primer in (B) so as to identify the base located at the locus of the nucleic acid.

15 In another method of the invention, a biological sample containing a plurality of nucleic acid molecules is analyzed for the presence of a mutation or polymorphism at a locus of each nucleic acid molecule, for each nucleic acid molecule. This method includes steps of:

- (A) amplifying the nucleic acid molecule in the presence of a first primer having a 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements

- 30 -

of the invention to form an amplified molecule with a 5'-end with a sequence complementary to the sequence of the tag;

- (B) extending the amplified molecule in the presence of a polymerase and a second primer having 5'-end complementary the 3'-end of the amplified sequence, the 3'-end of the second primer extending to immediately adjacent said locus, in the presence of a plurality of nucleoside triphosphate derivatives each of which is: (i) capable of incorporation during transcription by the polymerase onto the 3'-end of a growing nucleotide strand; (ii) causes termination of polymerization; and (iii) capable of differential detection, one from the other, wherein there is a said derivative complementary to each possible nucleotide present at said locus of the amplified molecule;
- (C) specifically hybridizing the second primer to a tag complement having the tag complement sequence of (A); and
- (D) detecting the nucleotide derivative incorporated into the second primer in (B) so as to identify the base located at the locus of the nucleic acid;

wherein each tag of (A) is unique for each nucleic acid molecule and steps (A) and (B) are carried out with said nucleic molecules in the presence of each other.

Another method includes analyzing a biological sample that contains a plurality of double stranded complementary nucleic acid molecules for the presence of a mutation or polymorphism at a locus of each nucleic acid molecule, for each nucleic acid molecule. The method includes steps of:

- 5 (A) amplifying the double stranded molecule in the presence of a pair of first primers, each primer having an identical 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements of the invention to form amplified molecules with 5'-ends with a sequence complementary to the sequence of the tag;
- (B) extending the amplified molecules in the presence of a polymerase and a pair of second primers each second primer having a 5'-end complementary a 3'-end of the amplified sequence, the 3'-end of each said second primer extending to immediately adjacent said locus, in the presence of a plurality of nucleoside triphosphate derivatives each of which is: (i) capable of incorporation during transcription by the polymerase onto the 3'-end of a growing nucleotide strand; (ii) causes termination of polymerization; and (iii) capable of differential detection, one from the other;

- 31 -

- (C) specifically hybridizing each of the second primers to a tag complement having the tag complement sequence of (A); and
- (D) detecting the nucleotide derivative incorporated into the second primers in (B) so as to identify the base located at said locus; wherein the sequence of each tag of (A) is unique for each nucleic acid molecule and steps (A) and (B) are carried out with said nucleic molecules in the presence of each other.

In yet another aspect, the invention is a method of analyzing a biological sample containing a plurality of nucleic acid molecules for the presence of a mutation or polymorphism at a locus of each nucleic acid molecule, for each nucleic acid molecule, the method including

5 steps of:

- (a) hybridizing the molecule and a primer, the primer having a 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements of the invention and a 3'-end extending to immediately adjacent the locus;
- (b) enzymatically extending the 3'-end of the primer in the presence of a plurality of nucleoside triphosphate derivatives each of which is: (i) capable of enzymatic incorporation onto the 3'-end of a growing nucleotide strand; (ii) causes termination of said extension; and (iii) capable of differential detection, one from the other, wherein there is a said derivative complementary to each possible nucleotide present at said locus;
- (c) specifically hybridizing the extended primer formed in step (b) to a tag complement having the tag complement sequence of (a); and
- (d) detecting the nucleotide derivative incorporated into the primer in step (b) so as to identify the base located at the locus of the nucleic acid molecule;

wherein each tag of (a) is unique for each nucleic acid molecule and steps (a) and (b) are carried out with said nucleic molecules in the presence of each other.

The derivative can be a dideoxy nucleoside triphosphate.

Each respective complement can be attached as a uniform population of substantially identical complements in spacially discrete regions on one or more solid phase support(s).

10 Each tag complement can include a label, each such label being different for respective complements, and step (d) can include detecting the presence of the different labels for respective hybridization complexes of bound tags and tag complements.

- 32 -

Another aspect of the invention includes a method of determining the presence of a target suspected of being contained in a mixture.

The method includes the steps of:

- (i) labelling the target with a first label;
- (ii) providing a first detection moiety capable of specific binding to the target and including a first tag;
- (iii) exposing a sample of the mixture to the detection moiety under conditions suitable to permit (or cause) said specific binding of the molecule and target;
- (iv) providing a family of suitable tag complements of the invention wherein the family contains a first tag complement having a sequence complementary to that of the first tag;
- (v) exposing the sample to the family of tag complements under conditions suitable to permit (or cause) specific hybridization of the first tag and its tag complement;
- (vi) determining whether a said first detection moiety hybridized to a first said tag complement is bound to a said labelled target in order to determine the presence or absence of said target in the mixture.

Preferably, the first tag complement is linked to a solid support at a specific location of the support and step (vi) includes detecting the presence of the first label at said specified location.

Also, the first tag complement can include a second label and step (vi) includes detecting the presence of the first and second labels in a hybridized complex of the moiety and the first tag complement.

Further, the target can be selected from the group consisting of organic molecules, antigens, proteins, polypeptides, antibodies and nucleic acids. The target can be an antigen and the first molecule can be an antibody specific for that antigen.

The antigen is usually a polypeptide or protein and the labelling step can include conjugation of fluorescent molecules, digoxigenin, biotinylation and the like..

The target can be a nucleic acid and the labelling step can include incorporation of fluorescent molecules, radiolabelled nucleotide, digoxigenin, biotinylation and the like.

DETAILED DESCRIPTION OF THE INVENTION

FIGURES

Reference is made to the attached figures in which,

- 33 -

Figures 1A and 1B illustrate results obtained in the cross-hybridization experiments described in Example 1. Figure 1A shows the hybridization pattern found when a microarray containing all 100 probes (SEQ ID NOs:1 to 100) was hybridized with a 24mer oligonucleotide having the complementary sequence to SEQ ID NO:3 (target). Figure 1B shows the pattern observed when a similar array was hybridized with a mix of all 100 targets, i.e., oligonucleotides having the sequences complementary to SEQ ID NOs:1 to 100.

Figure 2 shows the intensity of the signal (MFI) for each perfectly matched sequence (indicated in Table I) and its complement obtained as described in Example 2.

Figure 3 is a three dimensional representation showing cross-hybridization observed for the sequences of Figure 2 as described in Example 2. The results shown in Figure 2 are reproduced along the diagonal of the drawing.

Figure 4 is illustrative of results obtained for an individual target (SEQ ID NO:23, target No. 16) when exposed to the 100 probes of Example 2. The MFI for each bead is plotted.

20 DETAILED EMBODIMENTS

The invention provides a family of minimally cross-hybridizing sequences. The invention includes a method for sorting complex mixtures of molecules by the use of families of the sequences as oligonucleotide sequence tags. The families of oligonucleotide sequence tags are designed so as to provide minimal cross hybridization during the sorting process. Thus any sequence within a family of sequences will not cross hybridize with any other sequence derived from that family under appropriate hybridization conditions known by those skilled in the art. The invention is particularly useful in highly parallel processing of analytes.

30

Families of Oligonucleotide Sequence Tags

The present invention includes a family of 24mer polynucleotides, that have been demonstrated to be minimally cross-hybridizing with each other. This family of polynucleotides is thus useful as a family of tags, and their complements as tag complements.

The oligonucleotide sequences that belong to families of sequences that do not exhibit cross hybridization behavior can be derived by computer programs (described in international patent publication NO. WO 01/59151). The programs use a method of generating a maximum number of minimally cross-hybridizing polynucleotide sequences that can be summarized as follows.

First, a set of sequences of a given length are created based on a given number of block elements. Thus, if a family of polynucleotide sequences 24 nucleotides (24mer) in length is desired from a set of 6 block elements, each element comprising 4 nucleotides, then a family of 24mers is generated considering all positions of the 6 block elements. In this case, there will be 6^6 (46,656) ways of assembling the 6 block elements to generate all possible polynucleotide sequences 24 nucleotides in length.

Constraints are imposed on the sequences and are expressed as a set of rules on the identities of the blocks such that homology between any two sequences will not exceed the degree of homology desired between these two sequences. All polynucleotide sequences generated which obey the rules are saved. Sequence comparisons are performed in order to generate an incidence matrix. The incidence matrix is presented as a simple graph and the sequences with the desired property of being minimally cross hybridizing are found from a clique of the simple graph, which may have multiple cliques. Once a clique containing a suitably large number of sequences is found, the sequences are experimentally tested to determine if it is a set of minimally cross hybridizing sequences. This method has been used to obtain the 100 non cross-hybridizing tags of Table I that are the subject of this patent application.

The method includes a rational approach to the selection of groups of sequences that are used to describe the blocks. For example there are n^4 different tetramers that can be obtained from n different nucleotides, non-standard bases or analogues thereof. In a more preferred embodiment there are 4^4 or 256 possible tetramers when natural nucleotides are used. More preferably 81 possible tetramers when only 3 bases are used A, T and G. Most preferably 32 different tetramers when all sequences have only one G.

Block sequences can be composed of a subset of natural bases most preferably A, T and G. Sequences derived from blocks that are deficient in one base possess useful characteristics, for example, in reducing potential secondary structure formation or reduced potential for cross hybridization with nucleic acids in nature. Sets of block sequences that are most preferable in constructing families of non cross hybridizing tag sequences should contribute approximately equivalent stability to the formation of the correct duplex as all other block sequences of the set. This should provide tag sequences that behave isothermally. This can be achieved for example by maintaining a constant base composition for all block sequences such as one G and three A's or T's for each block sequence. Preferably, non-cross hybridizing sets of block sequences will be comprised from blocks of sequences that are isothermal. The block sequences should be different from each other by at least one mismatch. Guidance for selecting such sequences

is provided by methods for selecting primer and or probe sequences that can be found in published techniques (Robertson et al., Methods Mol Biol;98:121-54 (1998); Rychlik et al, Nucleic Acids Research, 17:8543-8551 (1989); Breslauer et al., Proc Natl Acad Sci., 83:3746-3750 (1986)) and the like.

5 Additional sets of sequences can be designed by extrapolating on the original family of non cross hybridizing sequences by simple methods known to those skilled in the art.

A preferred family of 100 tags is shown as SEQ ID NOS:1 to 100 in Table I. Characterization of the family of 100 sequence tags was performed to determine the ability of these sequences to form specific duplex structures with their complementary sequences and to assess the potential for cross hybridization. The 100 sequences were synthesized and spotted onto glass slides where they were coupled to the surface by amine linkage. Complementary tag sequences were Cy3-labeled and hybridized individually to the array containing the family of 100 sequence tags. Formation of duplex structures was detected and quantified for each of the positions on the array. Each of the tag sequences performed as expected, that is the perfect match duplex was formed in the absence of significant cross hybridization under stringent hybridization conditions. The results of a sample hybridization are shown in Figure 1. Figure 1a shows the hybridization pattern seen when a microarray containing all 100 probes was hybridized with the target complementary to probe 181234. The 4 sets of paired spots correspond to the probe complementary to the target. Figure 1b shows the pattern seen when a similar array was hybridized with a mix of all 100 targets. These results indicate that the family of sequences which is the subject of this patent can be used as a family of non-cross hybridizing (tag) sequences.

The family of 100 non-cross-hybridizing sequences can be expanded by incorporating additional tetramer sequences that are used in constructing further 24mer oligonucleotides. In one example, four additional words were included in the generation of new sequences to be considered for inclusion as non-cross talkers in a family of sequences that were obtained from the above method using 10 tetramers. In this case, the four additional words were selected to avoid potential homologies with all potential combinations of other words: YYXW (TTAG); WYYX (GTTA); XYXW (ATAG) and WYYY (GTTT). The total number of sequences containing six words using the 14 possible words is 14^6 or 7,529,536. These sequences were screened to eliminate sequences that contain repetitive regions that present potential hybridization problems such as four or more of a similar base (e.g., AAAA or TTTT) or pairs of G's. Each 40 of these sequences was compared to the sequence set of the original family of

- 36 -

100 non-cross-hybridizing sequences (SEQ ID NOS:1 to 100). Any new sequence that contained a minimal threshold of homology (that does not include the use of insertions or deletions) such as 15 or more matches with any of the original family of sequences was eliminated. In other words, if it was possible to align a new sequence with one or more of the original 100 sequences so as to obtain a maximum simple homology of 15/24 or more, the new sequence was dropped. "Simple homology" between a pair of sequences is defined here as the number of pairs of nucleotides that are matching (are the same as each other) in a comparison of two aligned sequences divided by the total number of potential matches. "Maximum simple homology" is obtained when two sequences are aligned with each other so as to have the maximum number of paired matching nucleotides. In any event, the set of new sequences so obtained was referred to as the "candidate sequences". One of the candidate sequences was arbitrarily chosen and referred to as sequence 101. All the candidate sequences were checked against sequence 101, and sequences that contained 15 or more non-consecutive matches (i.e., a maximum simple homology of 15/24 (62.5%) or more were eliminated. This results in a smaller set of candidate sequences from which another sequence is selected that is now referred to as sequence 102. The smaller set of candidate sequences is now compared to sequence 102 eliminating sequences that contained 15 or more non-consecutive matches and the process is repeated until there are no candidate sequences remaining. Also, any sequence selected from the candidate sequences is eliminated if it has 13 or more consecutive matches with any other previously selected candidate sequence.

The additional set of 73 tag sequences so obtained (SEQ ID NOS:101 to 173) is composed of sequences that when compared to any of SEQ ID NOS:1 to 100 of Table I have no greater similarity than the sequences of the original 100 sequence tags of Table I. The sequence set as derived from the original family of non cross hybridizing sequences, SEQ ID NOS:1 to 173, are expected to behave with similar hybridization properties to the sequences having SEQ ID NOS:1 to 100 since it is understood that sequence similarity correlates directly with cross hybridization (Southern et al., Nat. Genet.; 21, 5-9: 1999).

The set of 173 24mer oligonucleotides were expanded to include those having SEQ ID NOS:174 to 210 as follows. The 4mers WXYW, XYXW, WXXW, WYYW, XYYX, YXYX, YXXX and XYXY where W=G, X=A, and Y=U/T were used in combination with the fourteen 4mers used in the generation of SEQ ID NOS:1 to 173 to generate potential 24-base oligonucleotides. Excluded from the set were those containing the sequence patterns GG, AAAA and TTTT. To be included in the set of additional 24mers, a

- 37 -

sequence also had to have at least one of the 4mers containing two G's: WXYW (GATG), WYXW (GTAG), WXXW (GAAG), WYYW (GTTG) while also containing exactly six G's. Also required for a 24mer to be included was that there be at most six bases between every neighboring pair of 5 G's. Another way of putting this is that there are at most six non-G's between any two G's. Also, each G nearest the 5'-end of its oligonucleotide (the left-hand side as written in Table I) was required to occupy one of the first to seventh positions (counting the 5'- terminal position as the first position.) A set of candidate sequences 10 was obtained by eliminating any new sequence that was found to have a maximum simple homology of 16/24 or more with any of the previous set of 173 oligonucleotides (SEQ ID NOS:1 to 173). As above, an arbitrary 174th sequence was chosen and candidate sequences eliminated by comparison therewith. In this case the permitted maximum degree of 15 simple homology was 16/24. A second sequence was also eliminated if there were ten consecutive matches between the two (i.e., it was notionally possible to generate a phantom sequence containing a sequence of 10 bases that is identical to a sequence in each of the sequences being compared). A second sequence was also eliminated if it 20 was possible to generate a phantom sequence 20 bases in length or greater.

A property of the polynucleotide sequences shown in Table I is that the maximum block homology between any two sequences is never greater than 66 2/3 percent. This is because the computer algorithm by which the sequences were 25 initially generated was designed to prevent such an occurrence. It is within the capability of a person skilled in the art, given the family of sequences of Table I, to modify the sequences, or add other sequences while largely retaining the property of minimal-cross hybridization which the polynucleotides of Table I have been demonstrated to have.

30 There are 210 polynucleotide sequences given in Table I. Since all 210 of this family of polynucleotides can work with each other as a minimally cross-hybridizing set, then any plurality of polynucleotides that is a subset of the 210 can also act as a minimally cross-hybridizing set of polynucleotides. An application in which, for example, 30 molecules are to 35 be sorted using a family of polynucleotide tags and tag complements could thus use any group of 30 sequences shown in Table I. This is not to say that some subsets may be found in practical sense to be more preferred than others. For example, it may be found that a particular subset is more tolerant of a wider variety of conditions under which hybridization is 40 conducted before the degree of cross-hybridization becomes unacceptable.

- 38 -

It may be desirable to use polynucleotides that are shorter in length than the 24 bases of those in Table I. A family of subsequences (i.e., subframes of the sequences illustrated) based on those contained in Table I having as few as 10 bases per sequence could be chosen, so long as the 5 subsequences are chosen to retain homological properties between any two of the sequences of the family important to their non cross-hybridization.

The selection of sequences using this approach would be amenable to a computerized process. Thus for example, a string of 10 contiguous bases of the first 24mer of Table II could be selected: GATTGTATTGATTGAGATTAAAG.

A string of contiguous bases from the second 24mer could then be 10 selected and compared for maximum homology against the first chosen sequence:
TGATTGTAGTATGTATTGATAAAG

Systematic pairwise comparison could then be carried out to determine if the maximum homology requirement of 66 2/3 percent is violated:

15

Alignment	Matches
GATTGTATT	1
ATTGATAAAG	0
GATTGTATT	
ATTGATAAAG	1
GATTGTATT	
ATTGATAAAG	1
GATTGTATT	
ATTGATAAAG	1
GATTGTATT	
ATTGATAAAG	1
GATTGTATT	
ATTGATAAAG	3
GATTGTATT	
ATTGATAAAG	1
GATTGTATT	
ATTGATAAAG	2
GATTGTATT	
ATTGATAAAG	2
GATTGTATT	
ATTGATAAAG	5 (*)
GATTGTATT	
ATTGATAAAG	3
GATTGTATT	
ATTGATAAAG	3
GATTGTATT	
ATTGATAAAG	2
GATTGTATT	
ATTGATAAAG	1
GATTGTATT	
ATTGATAAAG	1
GATTGTATT	
ATTGATAAAG	3
GATTGTATT	

- 39 -

GATTTGTATT	1
ATTGATAAAG	
GATTTGTATT	0
ATTGATAAAG	

As can be seen, the maximum homology between the two selected subsequences is 50 percent (5 matches out of the total length of 10), and so these two sequences are compatible with each other.

5 A 10mer subsequence can be selected from the third 24mer sequence of Table I, and pairwise compared to each of the first two 10mer sequences to determine its compatibility therewith, etc. and in this way a family of 10mer sequences developed.

10 It is within the scope of this invention, to obtain families of sequences containing 11mer, 12mer, 13mer, 14mer, 15mer, 16mer, 17mer, 18mer, 19mer, 20mer, 21mer, 22mer and 23mer sequences by analogy to that shown for 10mer sequences.

15 It may be desirable to have a family of sequences in which there are sequences greater in length than the 24mer sequences shown in Table I. It is within the capability of a person skilled in the art, given the family of sequences shown in Table I, to obtain such a family of sequences. One possible approach would be to insert into each sequence at one or more locations a nucleotide, non natural base or analogue such that the longer sequence should not have greater similarity than any two of the original non 20 cross hybridizing sequences of Table I and the addition of extra bases to the tag sequences should not result in a major change in the thermodynamic properties of the tag sequences of that set for example the GC content must be maintained between 10%-40% with a variance from the average of 20%. This method of inserting bases could be used to obtain a family of sequences up to 25 40 bases long.

Given a particular family of sequences that can be used as a family of tags (or tag complements), e.g., those of Table I or Table II, or the combined sequences of these two tables, a skilled person will readily recognize variant families that work equally as well.

30 Again taking the sequences of Table I for example, every T could be converted to an A and vice versa and no significant change in the cross-hybridization properties would be expected to be observed. This would also be true if every G were converted to a C.

35 Also, all of the sequences of a family could be taken to be constructed in the 5'-3' direction, as is the convention, or all of the constructions of sequences could be in the opposite direction (3'-5').

- 40 -

There are additional modifications that can be carried out. For example, C has not been used in the family of sequences. Substitution of C in place of one or more G's of a particular sequence would yield a sequence that is at least as low in homology with every other sequence of the family as the particular sequence chosen to be modified was. It is thus possible to substitute C in place of one or more G's in any of the sequences shown in Table I. Analogously, substituting of C in place of one or more A's is possible, or substituting C in place of one or more T's is possible.

It is preferred that the sequences of a given family are of the same, or roughly the same length. Preferably, all the sequences of a family of sequences of this invention have a length that is within five bases of the base-length of the average of the family. More preferably, all sequences are within four bases of the average base-length. Even more preferably, all or almost all sequences are within three bases of the average base-length of the family. Better still, all or almost all sequences have a length that is within one of the base-length of the average of the family.

It is also possible for a person skilled in the art to derive sets of sequences from the family of sequences that is the subject of this patent and remove sequences that would be expected to have undesirable hybridization properties.

Methods For Synthesis Of Oligonucleotide Families

Preferably oligonucleotide sequences of the invention are synthesized directly by standard phosphoramidite synthesis approaches and the like (Caruthers et al, Methods in Enzymology; 154, 287-313: 1987; Lipshutz et al, Nature Genet.; 21, 20-24: 1999; Fodor et al, Science; 251, 763-773: 1991). Alternative chemistries involving non natural bases such as peptide nucleic acids or modified nucleosides that offer advantages in duplex stability may also be used (Hacia et al; Nucleic Acids Res.; 27: 4034-4039, 1999; Nguyen et al, Nucleic Acids Res.; 27, 1492-1498: 1999; Weiler et al, Nucleic Acids Res.; 25, 2792-2799:1997). It is also possible to synthesize the oligonucleotide sequences of this invention with alternate nucleotide backbones such as phosphorothioate or phosphoroamidate nucleotides. Methods involving synthesis through the addition of blocks of sequence in a step wise manner may also be employed (Lyttle et al, Biotechniques, 19: 274-280 (1995). Synthesis may be carried out directly on the substrate to be used as a solid phase support for the application or the oligonucleotide can be cleaved from the support for use in solution or coupling to a second support.

Solid Phase Supports

There are several different solid phase supports that can be used with the invention. They include but are not limited to slides, plates, chips, membranes, beads, microparticles and the like. The solid phase supports can also vary in the materials that they are composed of including plastic, glass, silicon, nylon, polystyrene, silica gel, latex and the like. The surface of the support is coated with the complementary sequence of the same.

In preferred embodiments, the family of tag complement sequences are derivatized to allow binding to a solid support. Many methods of derivatizing a nucleic acid for binding to a solid support are known in the art (Hermanson G., *Bioconjugate Techniques*; Acad. Press: 1996). The sequence tag may be bound to a solid support through covalent or non-covalent bonds (Iannone et al, *Cytometry*; 39: 131-140, 2000; Matson et al, *Anal. Biochem.*; 224: 110-106, 1995; Proudnikov et al, *Anal Biochem*; 259: 34-41, 1998; Zammattéo et al, *Analytical Biochemistry*; 280:143-150, 2000). The sequence tag can be conveniently derivatized for binding to a solid support by incorporating modified nucleic acids in the terminal 5' or 3' locations.

A variety of moieties useful for binding to a solid support (e.g., biotin, antibodies, and the like), and methods for attaching them to nucleic acids, are known in the art. For example, an amine-modified nucleic acid base (available from, e.g., Glen Research) may be attached to a solid support (for example, Covalink-NH, a polystyrene surface grafted with secondary amino groups, available from Nunc) through a bifunctional crosslinker (e.g., bis(sulfosuccinimidyl suberate), available from Pierce). Additional spacing moieties can be added to reduce steric hindrance between the capture moiety and the surface of the solid support.

Attaching Tags to Analytes for Sorting

A family of oligonucleotide tag sequences can be conjugated to a population of analytes most preferably polynucleotide sequences in several different ways including but not limited to direct chemical synthesis, chemical coupling, ligation, amplification, and the like. Sequence tags that have been synthesized with primer sequences can be used for enzymatic extension of the primer on the target for example in PCR amplification.

Detection of Single Nucleotide Polymorphisms Using Primer Extension

There are a number of areas of genetic analysis where families of non-cross hybridizing sequences can be applied including disease diagnosis, single nucleotide polymorphism analysis, genotyping, expression analysis and the

like. One such approach for genetic analysis referred to as the primer extension method (also known as Genetic Bit Analysis (Nikiforov et al, Nucleic Acids Res.; 22, 4167-4175: 1994; Head et al Nucleic Acids Res.; 25, 5065-5071: 1997)) is an extremely accurate method for identification of the nucleotide located at a specific polymorphic site within genomic DNA. In standard primer extension reactions, a portion of genomic DNA containing a defined polymorphic site is amplified by PCR using primers that flank the polymorphic site. In order to identify which nucleotide is present at the polymorphic site, a third primer is synthesized such that the polymorphic position is located immediately 3' to the primer. A primer extension reaction is set up containing the amplified DNA, the primer for extension, up to 4 dideoxynucleoside triphosphates, each labelled with a different fluorescent dye and a DNA polymerase such as the Klenow subunit of DNA Polymerase 1. The use of dideoxy nucleotides ensure that a single base is added to the 3' end of the primer, a site corresponding to the polymorphic site. In this way the identity of the nucleotide present at a specific polymorphic site can be determined by the identity of the fluorescent dye-labelled nucleotide that is incorporated in each reaction. One major drawback to this approach is its low throughput. Each primer extension reaction is carried out independently in a separate tube.

Universal sequences can be used to enhance the throughput of primer extension assay as follows. A region of genomic DNA containing multiple polymorphic sites is amplified by PCR. Alternately, several genomic regions containing one or more polymorphic sites each are amplified together in a multiplexed PCR reaction. The primer extension reaction is carried out as described above except that the primers used are chimeric, each containing a unique universal tag at the 5' end and the sequence for extension at the 3' end. In this way, each gene-specific sequence would be associated with a specific universal sequence. The chimeric primers would be hybridized to the amplified DNA and primer extension carried out as described above. This would result in a mixed pool of extended primers, each with a specific fluorescent dye characteristic of the incorporated nucleotide. Following the primer extension reaction, the mixed extension reactions are hybridized to an array containing probes that are reverse complements of the universal sequences on the primers. This would segregate the products of a number of primer extension reactions into discrete spots. The fluorescent dye present at each spot would then identify the nucleotide incorporated at each specific location.

Kits Using Families Of Tag Sequences

The families of non cross-hybridizing sequences may be provided in kits for use in for example genetic analysis. Such kits include at least one set 5 of non cross hybridizing sequences in solution or on a solid support. Preferably the sequences are attached to microparticles and are provided with buffers and reagents that are appropriate for the application. Reagents may include enzymes, nucleotides, fluorescent labels and the like that would be required for specific applications. Instructions for correct use of the kit 10 for a given application will be provided.

EXAMPLES**EXAMPLE 1 - Demonstration of Non Cross Talk Behavior on Solid Array**

15 One hundred oligonucleotide probes corresponding to a family of non-cross talking oligonucleotides from Table I were synthesized by Integrated DNA Technologies (IDT, Coralville IA). These oligonucleotides incorporated a C₆ aminolink group coupled to the 5' end of the oligo through a C₁₈ ethylene glycol spacer. These probes were used to prepare microarrays as follows. The 20 probes were resuspended at a concentration of 50 μM in 150 mM NaPO₄, pH 8.5. The probes were spotted onto the surface of a SuperAldehyde slide (Telechem Int., Sunnyvale CA) using an SDDC-II microarray spotter (ESI, Toronto Ontario, Canada). The spots formed were approximately 120 μM in diameter with 200 μM centre-to-centre spacing. Each probe was spotted 8 times on each 25 microarray. Following spotting, the arrays were processed essentially as described by the slide manufacturer. Briefly, the arrays were treated with 67 mM sodium borohydride in PBS/EtOH (3:1) for 5 minutes then washed with 4 changes of 0.1% SDS. The arrays were not boiled.

30 One hundred labelled oligonucleotide targets were also synthesized by IDT. The sequence of these targets corresponded to the reverse complement of the 100 probe sequences. The targets were labelled at the 5' end with Cy3.

Each Cy3-labeled target oligonucleotide was hybridized separately to two microarrays each of which contained all 100 oligonucleotide probes. 35 Hybridizations were carried out at 42°C for 2 hours in a 40 μl reaction and contained 40 nM of the labelled target suspended in 10 mM TrisHCl, pH 8.3, 50 mM KCl, 0.1% Tween 20. These are low stringency hybridization conditions designed to provide a rigorous test of the performance of the family of non-cross hybridizing sequences. Hybridizations were carried out by depositing

- 44 -

the hybridization solution on a clean cover slip then carefully positioning the microarray slide over the cover slip in order to avoid bubbles. The slide was then inverted and transferred to a humid chamber for incubation. Following hybridization, the cover slip was removed and the microarray was washed in hybridization buffer for 15 minutes at room temperature. The slide was then dried by brief centrifugation.

Hybridized microarrays were scanned using a ScanArray Lite (GSI-Lumonics, Billerica MA). The laser power and photomultiplier tube voltage used for scanning each hybridized microarray were optimized in order to maximize the signal intensity from the spots representing the perfect match.

The results of a sample hybridization are shown in Figures 1A and 1B. Figure 1A shows the hybridization pattern seen when a microarray containing all 100 probes was hybridized with the target complementary to probe 181234. The 4 sets of paired spots correspond to the probe complementary to the target. Figure 1b shows the pattern seen when a similar array was hybridized with a mix of all 100 targets.

Similar results to those illustrated in Figure 1a were obtained for all of the sequences tested, and the feasibility of the use of molecules containing oligonucleotides containing SEQ ID NOS:1 to 100 as a set of tags (or tag complements) is thus established.

EXAMPLE 2 - Cross Talk Behavior of Sequence on Beads

A group of 100 of the sequences of Table I was tested for feasibility for use as a family of minimally cross-hybridizing oligonucleotides. The 100 sequences selected are separately indicated in Table I along with the numbers assigned to the sequences in the tests.

The tests were conducted using the Luminex LabMAP™ platform available from Luminex Corporation, Austin, Texas, U.S.A. The one hundred sequences, used as probes, were synthesized as oligonucleotides by Integrated DNA Technologies (IDT, Coralville, Iowa, U.S.A.). Each probe included a C₆ aminolink group coupled to the 5'-end of the oligonucleotide through a C₁₂ ethylene glycol spacer. The C₆ aminolink molecule is a six carbon spacer containing an amine group that can be used for attaching the oligonucleotide to a solid support. One hundred oligonucleotide targets (probe complements), the sequence of each being the reverse complement of the 100 probe sequences, were also synthesized by IDT. Each target was labelled at its 5'-end with biotin. All oligonucleotides were purified using standard desalting procedures, and were reconstituted to a concentration of approximately 200 μM in sterile, distilled water for use. Oligonucleotide concentrations were

determined spectrophotometrically using extinction coefficients provided by the supplier.

Each probe was coupled by its amino linking group to a carboxylated fluorescent microsphere of the LabMAP system according to the Luminex[™] protocol. The microsphere, or bead, for each probe sequence has unique, or spectrally distinct, light absorption characteristics which permits each probe to be distinguished from the other probes. Stock bead pellets were dispersed by sonication and then vortexing. For each bead population, approximately five million microspheres (400 µL) were removed from the stock tube using barrier tips and added to a 1.5 mL Eppendorf tube (USA Scientific). The microspheres were then centrifuged, the supernatant was removed, and beads were resuspended in 25 µL of 0.2 M MES (2-(N-morpholino)ethane sulfonic acid) (Sigma), pH 4.5, followed by vortexing and sonication. One nmol of each probe (in a 25 µL volume) was added to its corresponding bead population. A volume of 2.5 µL of EDC cross-linker (1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (Pierce), prepared immediately before use by adding 1.0 mL of sterile ddH₂O to 10 mg of EDC powder, was added to each microsphere population. Bead mixes were then incubated for 30 minutes at room temperature in the dark with periodic vortexing. A second 2.5 µL aliquot of freshly prepared EDC solution was then added followed by an additional 30 minute incubation in the dark. Following the second EDC incubation, 1.0 mL of 0.02% Tween-20 (BioShop) was added to each bead mix and vortexed. The microspheres were centrifuged, the supernatant was removed, and the beads were resuspended in 1.0 mL of 0.1% sodium dodecyl sulfate (Sigma). The beads were centrifuged again and the supernatant removed. The coupled beads were resuspended in 100 µL of 0.1 M MES pH 4.5. Bead concentrations were then determined by diluting each preparation 100-fold in ddH₂O and enumerating using a Neubauer BrightLine Hemacytometer. Coupled beads were stored as individual populations at 2-8°C protected from light.

The relative oligonucleotide probe density on each bead population was assessed by Terminal Deoxynucleotidyl Transferase (TdT) end-labelling with biotin-ddUTPs. TdT was used to label the 3'-ends of single-stranded DNA with a labeled ddNTP. Briefly, 180 µL of the pool of 100 bead populations (equivalent to about 4000 of each bead type) to be used for hybridizations was pipetted into an Eppendorf tube and centrifuged. The supernatant was removed, and the beads were washed in

- 46 -

1x TdT buffer. The beads were then incubated with a labelling reaction mixture, which consisted of 5x TdT buffer, 25mM CoCl₂, and 1000 pmol of biotin-16-ddUTP (all reagents were purchased from Roche). The total reaction volume was brought up to 85.5 μ L with sterile, distilled H₂O, 5 and the samples were incubated in the dark for 1 hour at 37°C. A second aliquot of enzyme was added, followed by a second 1 hour incubation. Samples were run in duplicate, as was the negative control, which contained all components except the TdT. In order to remove unincorporated biotin-ddUTP, the beads were washed 3 times with 10 200 μ L of hybridization buffer, and the beads were resuspended in 50 μ L of hybridization buffer following the final wash. The biotin label was detected spectrophotometrically using SA-PE (streptavidin-phycoerythrin conjugate). The streptavidin binds to biotin and the phycoerythrin is spectrally distinct from the probe beads. The 10mg/mL stock of SA-PE 15 was diluted 100-fold in hybridization buffer, and 15 μ L of the diluted SA-PE was added directly to each reaction and incubated for 15 minutes at 37°Celsius. The reactions were analyzed on the Luminex¹⁰⁰ LabMAP. Acquisition parameters were set to measure 100 events per bead using a 20 sample volume of 50 μ L.

The results obtained are shown in Figure 2. As can be seen the Mean Fluorescent Intensity (MFI) of the beads varies from 277.75 to 2291.08, a range of 8.25 -fold. Assuming that the labelling reactions are complete for all of the oligonucleotides, this illustrates the signal intensity that would be obtained for each type of bead at this concentration if the target (i.e., labelled complement) was bound to 25 the probe sequence to the full extent possible.

The cross-hybridization of targets to probes was evaluated as follows. 100 oligonucleotide probes linked to 100 different bead populations, as described above, were combined to generate a master bead mix, enabling multiplexed reactions to be carried out. The pool 30 of microsphere-immobilized probes was then hybridized individually with each biotinylated target. Thus, each target was examined individually for its specific hybridization with its complementary bead-immobilized sequence, as well as for its non-specific hybridization with the other 35 99 bead-immobilized universal sequences present in the reaction. For each hybridization reaction, 25 μ L bead mix (containing about 2500 of each bead population in hybridization buffer) was added to each well of a 96-well Thermowell PCR plate and equilibrated at 37°C. Each target was diluted to a final concentration of 0.002 fmol/ μ L in hybridization

- 47 -

buffer; and 25 μ L (50 fmol) was added to each well, giving a final reaction volume of 50 μ L. Hybridization buffer consisted of 0.2 M NaCl, 0.1 M Tris, 0.08% Triton X-100, pH 8.0 and hybridizations were performed at 37°C for 30 minutes. Each target was analyzed in triplicate and six background samples (i.e. no target) were included in each plate. A SA-PE conjugate was used as a reporter, as described above. The 10 mg/mL stock of SA-PE was diluted 100-fold in hybridization buffer, and 15 μ L of the diluted SA-PE was added directly to each reaction, without removal of unbound target, and incubated for 15 minutes at 37°C. Finally, an additional 35 μ L of hybridization buffer was added to each well, resulting in a final volume of 100 μ L per well prior to analysis on the Luminex[™] LabMAP. Acquisition parameters were set to measure 100 events per bead using a sample volume of 80 μ L.

The percent hybridization was calculated for any event in which the NET MFI was at least 3 times the zero target background. In other words, a calculation was made for any sample where $(MFI_{sample} - MFI_{zero\ target\ background}) / MFI_{zero\ target\ background} \geq 3$.

A "positive" cross-talk event (i.e., significant mismatch or cross-hybridization) was defined as any event in which the net median fluorescent intensity ($MFI_{sample} - MFI_{zero\ target\ background}$) generated by a mismatched hybrid was greater than or equal to the arbitrarily set limit of 10% that of the perfectly matched hybrid determined under identical conditions. As there are 100 probes and 100 targets, there are $100 \times 100 = 10,000$ possible different interactions possible of which 100 are the result of perfect hybridizations. The remaining 9900 result from hybridization of a target with a mismatched probe.

The results obtained are illustrated in Figure 3. The ability of each target to be specifically recognized by its matching probe is shown. Of the possible 9900 non-specific hybridization events that could have occurred when the 100 targets were each exposed to the pool of 100 probes, 6 events were observed. Of these 6 events, the highest non-specific event generated a signal equivalent to 10.2 % of the signal observed for the perfectly matched pair (i.e. specific hybridization event).

Each of the 100 targets was thus examined individually for specific hybridization with its complement sequence as incorporated onto a microsphere, as well as for non-specific hybridization with the complements of the other 99 target sequences. Representative

- 48 -

hybridization results for target 16 (complement of probe 16, Table I) are shown in Figure 4. Probe 16 was found to hybridize only to its perfectly-matched target. No cross-hybridization with any of the other 99 targets was observed.

5 The foregoing results demonstrate the possibility of incorporating the 210 sequences of Table I, or any subset thereof, into a multiplexed system with the expectation that most if not all sequences can be distinguished from the others by hybridization. That is, it is possible to distinguish each target from the other targets by
10 hybridization of the target with its precise complement and minimal hybridization with complements of the other targets.

EXAMPLE 3 - Tag Sequences used in Sorting Polynucleotides

15 The family of non cross hybridizing sequence tags or a subset thereof can be attached to oligonucleotide probe sequences during synthesis and used to generate amplified probe sequences. In order to test the feasibility of PCR amplification with non cross hybridizing sequence tags and subsequently addressing each respective sequence to its appropriate location on two-dimensional or bead arrays, the following experiment was devised. A 24mer
20 tag sequence was connected in a 5'-3' specific manner to a p53 exon specific sequence (20mer reverse primer). The connecting p53 sequence represented the inverse complement of the nucleotide gene sequence. To facilitate the subsequent generation of single stranded DNA post-amplification the tag-
25 Reverse primer was synthesized with a phosphate modification (PO_4) on the 5'-end. A second PCR primer was also generated for each desired exon, which represented the Forward (5'-3') amplification primer. In this instance the Forward primer was labeled with a 5'-biotin modification to allow detection with Cy3-avidin or equivalent.

A practical example of the aforementioned description is as follows:
30 For exon 1 of the human p53 tumor suppressor gene sequence the following tag-Reverse primer was generated:

222087

222063

5'-PO4-GATTGTAAGATTTGATAAAAGTGT-A-TCCAGGGAAAGCGTGTACCGTCGT-3'

35 Tag Sequence # 3

Exon 1 Reverse

The numbering above the Exon-1 reverse primer represents the genomic nucleotide positions of the indicated bases.

The corresponding Exon-1 Forward primer sequence is as follows:

- 49 -

221873

221896

5' -Biotin-TCATGGCGACTGTCCAGCTTGTG-3'

In combination these primers will amplify a product of 214 bp plus a
5 24 bp tag extension yielding a total size of 238 bp.
Once amplified, the PCR product was purified using a QIAquick PCR
purification kit and the resulting DNA was quantified. To generate
single stranded DNA, the DNA was subjected to λ -exonuclease digestion
thereby resulting in the exposure of a single stranded sequence (anti-
10 tag) complementary to the tag-sequence covalently attached to the solid
phase array. The resulting product was heated to 95°C for 5 minutes
and then directly applied to the array at a concentration of 10-50 nM.
Following hybridization and concurrent sorting, the tag-Exon 1
15 sequences were visualized using Cy3-streptavidin. In addition to
direct visualization of the biotinylated product, the product itself
can now act as a substrate for further analysis of the amplified
region, such as SNP detection and haplotype determination.

A number of additional methods for the detection of single
nucleotide polymorphisms, including but not limited to, allele specific
20 polymerase chain reaction (ASPCR), allele specific primer extension
(ASPE) and oligonucleotide ligation assay (OLA) can be performed by
those skilled in the art in combination with the tag sequences
described herein.

25 DEFINITIONS

Non cross hybridization: Describes the absence of hybridization
between two sequences that are not perfect complements of each other.

Cross Hybridization: The hydrogen bonding of a single-stranded
DNA sequence that is partially but not entirely complementary to a
30 single-stranded substrate.

Homology: How closely related two or more separate strands of DNA
are to each other, based on their base sequences.

Analogue: A chemical which resembles a nucleotide base. A base
which does not normally appear in DNA but can substitute for the ones
35 which do, despite minor differences in structure.

Complement: The opposite or "mirror" image of a DNA sequence. A
complementary DNA sequence has an "A" for every "T" and a "C" for every
"G". Two complementary strands of single stranded DNA, for example a
tag sequence and it's complement, will join to form a double-stranded
40 molecule.

- 50 -

Complementary DNA (cDNA): DNA that is synthesized from a messenger RNA template; the single-stranded form is often used as a probe in physical mapping.

Oligonucleotide: Refers to a short nucleotide polymer whereby the 5 nucleotides may be natural nucleotide bases or analogues thereof.

Tag: Refers to an oligonucleotide that can be used for specifically sorting analytes with at least one other oligonucleotide that when used together do not cross hybridize.

Similar Homology: In the context of this invention, pairs of sequences 10 are compared with each other based on the amount of "homology" between the sequences. By way of example, two sequences are said to have a 50% "maximum homology" with each other if, when the two sequences are aligned side-by-side with each other so to obtain the (absolute) maximum number of identically paired bases, the number of identically paired bases is 50% of the total 15 number of bases in one of the sequences. (If the sequences being compared are of different lengths, then it would be of the total number of bases in the shorter of the two sequences.) Examples of determining maximum homology are as follows:

20 EXAMPLE 4 - Determining Maximum Homology

* *

A-A-B-B-C-C

B-D-C-D-D-D (2 out of 4 paired bases are the same)

25

* *

A-A-B-B-C-C

B-D-C-D-D-D (2 out of 3 paired bases are the same)

30

In this case, the maximum number of identically paired bases is two and there are two possible alignments yielding this maximum number. The total number of possible pairings is six giving 33 1/3 % (2/6) homology. The maximum amount of homology between the two sequences is thus 1/3.

35 EXAMPLE 5 - Determining Maximum Homology

* * *

A-A-B-B-C-A

A-A-D-D-C-D (3 out of 6 paired bases are the same)

40

- 51 -

In this alignment, the number of identically paired bases is three and the total number of possibly paired bases is six, so the homology between the two sequences is 3/6 (50%).

5

*

A-A-B-B-C-A

A-A-D-D-C-D (1 out of 1 paired bases are the same)

In this alignment, the number of identically paired bases is 1, so the
10 homology between the two sequences is 1/6 (16 2/3%).

The maximum homology between these two sequences is thus 50%.

Block sequence: Refers to a symbolic representation of a sequence of blocks. In its most general form a block sequence is a representative
15 sequence in which no particular value, mathematical variable, or other designation is assigned to each block of the sequence.

Incidence Matrix: As used herein is a well-defined term in the field of Discrete Mathematics. However, an incidence matrix cannot be defined without first defining a "graph". In the method described herein a subset of general
20 graphs called simple graphs is used. Members of this subcategory are further defined as follows.

A simple graph G is a pair (V, E) where V represents the set of vertices of the simple graph and E is a set of un-oriented edges of the simple graph. An edge is defined as a 2-component combination of members of
25 the set of vertices. In other words, in a simple graph G there are some pairs of vertices that are connected by an edge. In our application a graph is based on nucleic acid sequences generated using sequence templates and vertices represent DNA sequences and edges represent a relative property of any pair of sequences.

The incidence matrix is a mathematical object that allows one to describe any given graph. For the subset of simple graphs used herein, the simple graph $G=(V,E)$, and for a pre-selected and fixed ordering of vertices, $V=\{v_1, v_2, \dots, v_n\}$, elements of the incidence matrix $A(G) = [a_{ij}]$ are defined by the following rules:

35

- (1) $a_{ij}=1$ for any pair of vertices $\{v_i, v_j\}$ that is a member of the set of edges; and
- (2) $a_{ij}=0$ for any pair of vertices $\{v_i, v_j\}$ that is not a member of the set of edges.

40

This is an exact unequivocal definition of the incidence matrix. In effect, one selects the indices: 1, 2, ..., n of the vertices and then forms an (n x n) square matrix with elements $a_{ij}=1$ if the vertices v_i and v_j are connected by an edge and $a_{ij}=0$ if the vertices v_i and v_j are not connected by an edge.

To define the term "class property" as used herein, the term "complete simple graph" or "clique" must first be defined. The complete simple graph is required because all sequences that result from the method described herein should collectively share the relative property of any pair of sequences defining an edge of graph G, for example not violating the threshold rule that is, do not have a "maximum simple homology" greater than a predetermined amount, whatever pair of the sequences are chosen from the final set. It is possible that additional "local" rules, based on known or empirically determined behavior of particular nucleotides, or nucleotide sequences, are applied to sequence pairs in addition to the basic threshold rule.

In the language of a simple graph, $G=(V, E)$, this means in the final graph there should be no pair of vertices (no sequence pair) not connected by an edge (because an edge means that the sequences represented by v_i and v_j do not violate the threshold rule).

Because the incidence matrix of any simple graph can be generated by the above definition of its elements, the consequence of defining a simple complete graph is that the corresponding incidence matrix for a simple complete graph will have all off-diagonal elements equal to 1 and all diagonal elements equal to 0. This is because if one aligns a sequence with itself, the threshold rule is of course violated, and all other sequences are connected by an edge.

For any simple graph, there might be a complete subgraph. First, the definition of a subgraph of a graph is as follows. The subgraph $G_s=(V_s, E_s)$ of a simple graph $G=(V, E)$ is a simple graph that contains the subsets of vertices V_s of the set V of vertices and inclusion of the set V_s into the set V is immersion (a mathematical term). This means that one generates a subgraph $G_s=(V_s, E_s)$ of a simple graph G in two steps. First select some vertices V_s from G . Then select those edges E_s from G that connect the chosen vertices and do not select edges that connect selected with non selected vertices.

We desire a subgraph of G that is a complete simple graph. By using this property of the complete simple graph generated from the simple graph G of all sequences generated by the template based algorithm, the pairwise property of any pair of the sequences (violating/non-violating the

- 53 -

threshold rule) is converted into the property of all members of the set, termed "the class property".

By selecting a subgraph of a simple graph G that is a complete simple graph, this assures that, up to the tests involving the local rules described herein, there are no pairs of sequences in the resulting set that violate the threshold rule, also described above, independent of which pair of sequences in the set are chosen. This feature is called the "desired class property".

The present invention thus includes reducing the potential for non cross-hybridization behavior by taking into account local homologies of the sequences and appears to have greater rigor than known approaches. For example, the method described herein involves the sliding of one sequence relative to the other sequence in order to form a sequence alignment that would accommodate insertions or deletions. (Kane et al., Nucleic Acids Res.; 28, 4552-4557: 2000).

Table I
Sequence **No Assigned in Example 2(2)**

SEQ ID NO (1)	Sequence	No Assigned in Example 2(2)
1	GATTGTATTGATTGAGATTAAG	1
2	TGATTGTAGTATGTATTGATAAAAG	2
3	GATTGTAAGATTGATAAAAGTGT	3
4	GATTGAAGATTATTGGTAATGT	4
5	GATTGATTATTGTGATTGAAATTG	5
6	GATTGATTGTAAAAGATTGTTGA	6
7	ATTGGTAAATTGGTAAATGAATTG	7
8	ATTGGATTGATAAAGGTAAATGA	
9	GTAAGTAATGAATGTAAGGATT	8
10	GATTGATTGATTGATTGATTGAT	
11	TGATGATTAAAGAAAGTGATTGAT	
12	AAAGGATTGATTGATAAAAGTGAT	
13	TGTAGATTGTATGTATGTATGAT	10
14	GATTGATAAAAGAAAGGATTGATT	
15	GATTAAAGTGATTGATGATTGTA	11
16	AAAGAAAGAAAGAAAGAAAGTGAT	12
17	TGAAAAGGATTGATTGTATGTA	
18	AAAGTGTAGATTGATTAAAGAAAG	
19	AAAGTTGATTGATTGAAAAGGTAT	
20	TTGATTGAGATTGATTGAGTAT	
21	TGAATTGATGAATGAATGAAGTAT	15
22	GTAATGAAGTATGTATGTAAGTAA	
23	TGATGATTGAAATGAAGATTGATT	16
24	TGATAAAAGTGATAAAGGATTAAAG	17
25	TGATTGAGATTGAGATTGATTTGA	18
26	TGTAGTAAGATTGATTAAGGTA	
27	GTATAAAGGATTGATTGAAAAG	
28	GTATTTGAGTAAGTAATTGATTGA	19
29	GTAAAAAGTTGAGTATTGAAAAG	
30	GATTGATAAAGGATTGATTGAT	
31	GATTGTATTGAAAGTATTGAAAAG	
32	TGATGATTGATGAAAAGTTGA	20

- 54 -

Table I

SEQ ID NO (1)	Sequence	No Assigned in Example 2 (2)
33	TGATTTGAGATTAAAGAAAGGATT	21
34	TGATTGAATTGAGAAAAAGGATT	22
35	AAAGTGTAAAAGGATTGATGTAT	
36	AAAGGTATTTGAGATTGATTGAA	
37	AAAGTTGAGATTGAAATGATTGAA	23
38	TGTATTGAAAAGGTATGATTGAA	
39	GTATTGTATTGAAAAGGTATTGA	24
40	TTGAGTAATGATAAAGTGAAGATT	
41	TGAAGATTGAAAGTAATTGAAAAG	25
42	TGAAAAAGTGTAGATTGAGTAA	26
43	TGTATGAATGAAGATTGATTGTA	
44	AAAGTTGAGTATTGATTGAAAAG	27
45	GATTTGTAGATTGTATTGAGATT	
46	AAAGAAAAGGATTGTAGTAAGATT	29
47	GTAAAAAGAAAAGGTATAAAGGTTAA	30
48	GATTAAGTTGATTGAAAAGTGA	31
49	TGAAAAAGGTAATTGATGTATGAA	
50	AAAGGATTAAAGTGAAGTAATTGA	33
51	ATGAATTGGTATGTATATGAATGA	34
52	TGAAATGAATGAATGATGAAATTG	35
53	ATTGATTGTGAATGAAATGAATTG	36
54	ATTGAAAGATGAAAAGATGAAAAG	37
55	ATTGTTGAAAAGTGTAAATGATTGA	38
56	ATGATGTAATGAAAAGATTGTGTA	39
57	AAAGATTGAAAAGATGATGTAATTG	
58	ATTGATGAGTATATTGTGAGTAA	41
59	AAAGATTGTGTAATTGATGATGAA	
60	AAAGGTATATTGTGTAATGAGTAA	
61	TGTAATGAGTATTGTAATTGAAAG	43
62	GTATAAAAGAAAAGATTGGTAAATGA	44
63	TTGAGTAATTGAATTGTTGAAATGA	45
64	TGTATTGAATGAATTGTTGATGTA	46
65	TGTAATTGGTAAATGAGTAAAAG	
66	TGAATGAAAATTGATGAGTATAAAG	
67	GTAAGTAAATTGAAAGATTGATGA	49
68	GTAATGATGATATTGGTATATTG	50
69	ATTGTTGATGATTGATTGAAATGA	51
70	ATTGTGAAGTATAAAGATGATTGA	52
71	ATGAAAAGTTGAGTAAATTGTGAT	
72	ATGAATTGAAAAGTATTGAAAAG	54
73	GTAATTGATGAAAAGTTGATGAT	
74	AAAGTGATGTATATGAGTAAATTG	56
75	GTAATGATAAAGATGATGATATTG	57
76	TTGAAAAGATTGGTAATGATATGA	
77	AAAGTAAAAAGATTGATTGATGA	59
78	ATTGATGAGATTGATTATTGTGTA	
79	ATGAGATTATTGGATTGAGATT	60
80	TGAAGATTATGAATTGGTAAGATT	61
81	ATTGGATTATGAGATTATGATTGA	62
82	ATTGTTGAATTGGATTAAAGATGA	
83	AAAGATGAGTAAGTAAATTGGATT	
84	AAAGGTAAAGATTATTGATGAAAAG	65
85	ATTGATGAGATTAAAGTTGATGATTG	
86	GATTATTGGATTATGAAAAGGATT	
87	GATTTGTAATTGTTGAGTAAATGA	67
88	AAAGAAAAGATTGTTGAGATTATGA	68
89	GTATAAAGGATTGGATTGATGA	

Table I

SEQ ID NO (1)	Sequence	No Assigned in Example 2 (2)
90	TTGAGATTGAAATGAATTGTGTA	
91	GTATATTGATTGTGTAATGAAAAG	
92	TGATATGAATTGGATTATTGGTAT	70
93	ATGAATGATGAATGATGATTATTG	
94	ATGAATTGATTGGATTGTAATGAT	71
95	GATTGTAATTGAGTAAATTGATGA	
96	GATTATTGGATTAAGGTAATGTA	72
97	ATTGTTGAATTGATGAGATTGAT	73
98	GATTATGAGTAAATTGATTGTGAT	
99	GATTATTGTTGATGAATGATATTG	
100	TGTAAAAGATTGAAAGGTATGATT	75
101	GTATTTAGATGAGTTGTTAGATT	76
102	TGAAGTTATGTAATAGAAAGTGAT	
103	GTATGTATTGTATGTAGTTAATTG	77
104	TGATATAGATAGTTAGATAGATAG	78
105	ATGATGATGATTGTAGTTATGAA	79
106	TTAGTGAATGTATTAGTTGATGTA	
107	GTTAGTTAGATTATTGTTAGTTAG	80
108	GTTAATTGTTAGTTGTTATGAA	
109	GTTATGAAATAGTGTATTGTTAG	
110	ATTGTTAGAAAGTGTAGATTAAAG	81
111	ATGAGTATGTTATTAGTGTATGTA	82
112	TGTAATAGTGAAGTTAGATTGTAT	83
113	ATTGATAGATGATTAGTTAGTTGA	84
114	ATGAGTTGTTATGAGATTAAAG	
115	TGATGTTGATTATGATGTAGTAT	85
116	ATGAGTTAGTTATGAAATTAGATGA	
117	ATTGTTAGTGTAGTTAGTAATTAG	86
118	TGATGTAAGTATTGATGTTAGTT	87
119	GATTGTAATAGAAAGTGAAGTAA	88
120	ATTGTGTATGAAAGTATTGTATGAT	
121	ATAGTGTATGTTATGAAAGATTGTA	
122	TTAGATGAAATTGTGAAGTATTAG	90
123	GTAAGTTATGATTGATGTTATGAA	91
124	GTATTGATGTTAAAGTGTAAAG	92
125	GATTGTAAGTAAGATTGTATATTG	
126	GTTTGTATTAGATGAAATAGAAAG	93
127	GTTTGTATTGTAATAGTGTAGTTA	
128	TGTATGTAGTATTAGAAAGATGA	
129	ATGAATTGATGATAAGAAAGTTAG	
130	TTAGTGTAGTAAGTTAAAGTGT	95
131	GTATGATTGTTGTAATTAGTGTAT	
132	GTTTAAAGTTAGTTGAGTTAGTAT	96
133	ATAGTGTATGTAGATTATGAGATT	97
134	TTGAATGATTAGTTGAGTATGATT	98
135	GTATGTAAGTTAGTATGATTGAA	
136	TGAGTATATTGTTGAAATTGTGAT	
137	ATAGTGTATGTATGTATGATAAG	
138	TTAGTGTATGTATATTGAAAG	
139	GTAAGATTATGAGTTATGATGTA	
140	GTTATGAAATTGTTAGTGTAGATT	99
141	GTTAGATTGTTAGTTAAAGATAG	100
142	TTAGTGTATGAAATGATGTAGATT	
143	AAAGTGTAGTTATTAGTTAGTTAG	
144	AAAGAAAGTGTATGATGTTATTAG	
145	GATTGTATATTGTGTATGATGATT	
146	TTGAGATTGTTATGATATGAGTAT	

- 56 -

Table I

SEQ ID NO(1)	Sequence	No Assigned in Example 2(2)
147	ATGAGTATGATTGTTATGATGTTT	
148	TGATTTAGTGAATTGTGTATTAG	
149	TGAATGTATGTAGTATGTTGTTA	
150	GTTAGTATTGTGATTATGAGTTA	
151	GTATATTGTGATTAGTTGAGATT	
152	GTTAGTTAAAGTTGAGATTGTTT	
153	GTATATTGTAGATGAGATTGTA	
154	TGATGTATGTTAGTTATGAATGA	
155	TGAGTATGTAATGTAGTATTGAA	
156	ATGAGTTATGTATTGAGTTAGTAT	
157	TGTATGATGATTATAGTTGAGTAA	
158	ATTGATGAATGAGTTGTTGATAAAG	
159	TTGAGTTATGATTAGAAAGAAAG	
160	TGATATTGTGAGTTAGTATTGAA	
161	ATAGAAAGTGAATGAGTATGTTA	
162	TTGATGTAGATTGTATGTATATAG	
163	TTGAGATTATAGTGTAGTTTATAG	
164	TGATGTTAGATTGTTGATTATTG	
165	TGTATTAGATAGTGATTGAAATGA	
166	GATTATGATGAATGTAGTATGTA	
167	TGAATGATTGATATGAATAGTGT	
168	GTAATGATTAGTGTATTGAGTTT	
169	TGTAGTAATGATTGATGATAAAG	
170	TGAAGATTGTTATTAGTGTATATTG	
171	GTATTTGAATGATGTAATAGTGT	
172	GTATATGATGTATTAGATTGAAAG	
173	AAAGTTAGATTGAAAGTGATAAAG	
174	GTAAGATGTTGATATAGAAGATTA	9
175	TAATATGAGATGAAAGTGAATTAG	
176	TTAGTGAAGAAGTATAGTTTATG	13
177	GTAGTTGAGAAGATAGTAATTAAT	
178	ATGAGATGATATTGAGAAGTAAT	
179	GATGTGAAGAAGATGAATATATAT	
180	AAAGTATAGTAAGATGTATAGTAG	14
181	GAAGTAATATGAGTAGTGTGAATAT	
182	TTGATAATGTTGTTGTTGTAG	28
183	TGAAGAAGAAGTATAATGATGAA	
184	GTAGATTAGTTGAAGTGAATAAT	32
185	TATAGTAGTGAAGATGATATATGA	
186	TATAATGAGTTGTTAGATATGTTG	
187	GTTGTGAAATTAGATGTGAAATAT	
188	TAATGTTGTAATAATGTAGAAAG	40
189	GTTTATAGTGAAAATATGAAGATAG	42
190	ATTATGAAGTAAGTTAATGAGAAG	47
191	GATGAAAGTAATGTTATTGTGAA	
192	ATTATTGAGATGTGAAGTTGTTT	48
193	TGTAGAAGATGAGATGTATAATTA	53
194	TAATTGAGTTGTTATAGTAG	
195	TGATATTAGTAAGAAGTTGAATAG	
196	GTTAGTTATTGAGAAGTGTATATA	55
197	GTAGTAATGTTAATGAATTAGTAG	58
198	GTTTGTGTTGATGTGAAATTGAAAT	
199	GTAAGTAGTAATTGAAATATGTAG	64
200	GTTTGAAGATATGTTGAAAGTATA	
201	ATGATAATTGAAAGATGTAAATGTT	
202	GTAGATAGTATAGTTGTAATGTTA	66
203	GATGTGAATGTAATATGTTATAG	69

- 57 -

Table I

SEQ ID NO(1)	Sequence	No Assigned in Example 2(2)
204	TGAAATTAGTTGTAAGATGTGA	74
205	TGTAGTATAAAGTATATGAAGTAG	63
206	ATATGTTGTTGAGTTGATAGTATA	89
207	ATTATTGAGTAGAAAGATAGAAAG	94
208	GTTGTTGAATATTGAATATAGTTG	
209	ATGAGAAGTTAGTAATGTAAATAG	
210	TGAAATGAGAAGATTAATGAGTT	

(1) Oligonucleotides having SEQ ID NOS:1 to 100 were used in experiments of Example 1.

(2) Oligonucleotides used in experiments of Example 2 are indicated in this column by the numbers assigned to them in the experiments.

All references referred to in this specification are incorporated herein by reference.

The scope of protection sought for the invention described herein is defined by the appended claims. It will also be understood that any elements recited above or in the claims, can be combined with the elements of any claim. In particular, elements of a dependent claim can be combined with any element of a claim from which it depends, or with any other compatible element of the invention.

This application claims priority from United States Provisional Patent Application Nos. 60/263,710 and 60/303,799, filed January 25, 2001 and July 10, 2001. Both of these documents are incorporated herein by reference.

- 58 -

Claims

1. A composition comprising molecules for use as tags or tag complements wherein each molecule comprises an oligonucleotide selected from a set of oligonucleotides based on a following group of sequences:

1	4	6	6	1	3
2	4	5	5	2	3
1	8	1	2	3	4
1	7	1	9	8	4
1	1	9	2	6	9
1	2	4	3	9	6
9	8	9	8	10	9
9	1	2	3	8	10
8	8	7	4	3	1
1	1	1	1	1	2
2	1	3	3	2	2
3	1	2	2	3	2
4	1	4	4	4	2
1	2	3	3	1	1
1	3	2	2	1	4
3	3	3	3	3	4
4	3	1	1	4	4
3	4	1	1	3	3
3	6	6	6	3	5
6	6	1	1	6	5
7	6	7	7	7	5
8	7	5	5	8	8
2	1	7	7	1	1
2	3	2	3	1	3
2	6	5	6	1	6
4	8	1	1	3	8
5	3	1	1	6	3
5	6	8	8	6	6
8	3	6	5	7	3
1	2	3	1	4	6
1	5	7	5	4	3
2	1	6	7	3	6
2	6	1	3	3	1
2	7	6	8	3	1
3	4	3	1	2	5
3	5	6	1	2	7
3	6	1	7	2	7
4	6	3	5	1	7
5	4	6	3	8	6
6	8	2	3	7	1
7	1	7	8	6	3
7	3	4	1	6	8
4	7	7	1	2	4
3	6	5	2	6	3
1	4	1	4	6	1
3	3	1	4	8	1
8	3	3	5	3	8
1	3	6	6	3	7
7	3	8	6	4	7
3	1	3	7	8	6
10	9	5	5	10	10

- 59 -

7	10	10	10	7	9
9	9	7	7	10	9
9	3	10	3	10	3
9	6	3	4	10	6
10	4	10	3	9	4
3	9	3	10	4	9
9	10	5	9	4	8
3	9	4	9	10	7
3	5	9	4	10	8
4	10	5	4	9	3
5	3	3	9	8	10
6	8	6	9	7	10
4	6	10	9	6	4
4	9	8	10	8	3
7	7	9	10	5	3
8	8	9	3	9	10
8	10	2	9	5	9
9	6	2	2	7	10
9	7	5	3	10	6
10	3	6	8	9	2
10	9	3	2	7	3
8	9	10	3	6	2
3	2	5	10	8	9
8	2	3	10	2	9
6	3	9	8	2	10
3	7	3	9	9	10
9	10	1	1	9	4
10	1	9	1	4	1
7	1	10	9	8	1
9	1	10	1	10	6
9	6	9	1	3	10
3	10	8	8	9	1
3	8	1	9	10	3
9	10	1	3	6	9
1	9	1	10	3	1
1	4	9	6	8	10
3	3	9	6	1	10
5	3	1	6	9	10
6	1	8	10	9	6
5	9	9	4	10	3
2	10	9	1	9	5
10	10	7	2	1	9
10	9	9	1	8	2
1	8	6	8	9	10
1	9	1	3	8	10
9	6	9	10	1	2
1	10	8	9	9	2
1	9	6	7	2	9
4	3	9	3	5	1
5	11	10	14	12	1
7	12	4	13	3	2
5	5	4	4	12	9
2	13	13	11	13	13
10	2	5	4	12	7
11	7	4	11	6	4
12	12	1	9	11	11
12	9	4	14	12	6
12	7	13	2	9	11
9	11	3	4	1	3
10	5	12	11	4	4

- 60 -

4	13	7	12	1	5
9	13	10	11	11	6
10	14	14	10	1	3
2	14	1	10	4	5
10	12	12	7	11	10
9	11	2	12	8	11
2	8	5	2	12	14
1	8	13	3	7	8
9	4	7	5	4	2
13	2	12	7	1	12
11	10	9	7	5	11
8	12	2	2	12	7
5	2	14	3	4	13
1	8	8	1	5	9
14	5	11	10	13	3
14	1	4	13	2	4
4	4	5	11	3	10
10	9	2	3	3	11
11	4	8	14	3	4
5	1	14	8	11	2
14	3	11	6	12	5
13	4	4	1	10	1
6	10	11	6	5	1
5	8	12	5	1	7
4	5	9	6	9	2
13	2	4	4	2	3
11	2	2	5	9	3
8	1	10	12	2	8
12	7	9	11	4	1
12	1	4	14	3	13
11	2	7	10	4	1
3	4	12	11	11	11
3	3	4	2	12	11
1	5	9	4	2	1
6	1	12	2	10	5
10	5	1	12	2	14
2	11	7	9	4	11
7	4	4	5	14	12
12	5	2	1	10	12
5	9	2	11	6	1
12	14	3	6	1	14
5	9	11	10	1	4
2	5	12	14	10	10
4	5	8	4	5	6
10	12	4	6	12	5
4	2	1	13	6	8
9	10	10	14	5	3
6	14	10	11	3	3
2	9	10	12	5	7
13	3	7	10	5	12
6	4	1	2	5	13
6	1	13	4	14	13
2	12	1	14	1	9
4	11	13	2	6	10
1	10	7	4	5	8
7	2	2	10	13	4
8	2	11	4	6	14
4	8	2	6	2	3
7	1	12	11	2	9
5	6	10	4	13	4

- 61 -

5	10	4	11	9	3
3	11	9	3	2	3
8	15	6	20	17	19
21	10	15	3	7	11
11	7	17	20	14	9
16	6	17	13	21	21
10	15	22	6	17	21
15	7	17	10	22	22
3	20	8	15	20	16
17	21	10	16	6	22
6	21	14	14	14	16
7	17	3	20	10	7
16	19	14	17	7	21
20	16	7	15	22	10
20	10	18	11	22	18
18	7	19	15	7	22
21	18	7	21	16	3
14	13	7	22	17	13
19	7	8	12	10	17
15	3	21	14	9	7
19	6	15	7	14	14
4	17	10	15	20	19
21	6	18	4	20	16
2	19	8	17	6	13
12	12	6	17	4	20
16	21	12	10	19	16
14	14	15	2	7	21
8	16	21	6	22	16
14	17	22	14	17	20
10	21	7	15	21	18
16	13	20	18	21	12
15	7	4	22	14	13
7	19	14	8	15	4
4	5	3	20	7	16
22	18	6	18	13	20
19	6	16	3	13	3
18	6	22	7	20	18
10	17	11	21	8	13
7	10	17	19	10	14

wherein:

- (A) each of 1 to 22 is a 4mer selected from the group of 4mers consisting of WWWW, WWWX, WWWY, WWXW, WWXX, WWXY, WYYW, WWYX, WWYY, WXWW, WXWX, WXWY, WXXW, WXXX, WXXY, WXYW, WXYX, WXYY, WYWW, WYWX, WYWy, WYXW, WYXX, WYXY, WYYW, WYYX, WYYY, XWWW, XWWX, XWWY, XWXW, XWXX, XWXY, XWYW, XWYY, XXWW, XXWX, XXWY, XXXW, XXXX, XXXY, XXYW, XXYX, XXYY, XYWW, XYWX, XYWY, XYYW, XYYY, YWWW, YWWX, YWWY, YWXW, YWXX, YWXY, YWYW, YWYY, YXWW, YXWX, YXWY, YXXW, YXXX, YXXY, YXYW, YXYX, YXYY, YYWW, YYWX, YYWY, YYXW, YYXX, YYXY, YYYW, YYYX, and YYYY, and
- (B) each of 1 to 22 is selected so as to be different from all of the others of 1 to 22;
- (C) each of W, X and Y is a base in which:

- 62 -

- (i) (a) W = one of A, T/U, G, and C,
X = one of A, T/U, G, and C,
Y = one of A, T/U, G, and C,
and each of W, X and Y is selected so as to be different
from all of the others of W, X and Y,
(b) an unselected said base of (i)(a) can be substituted any
number of times for any one of W, X and Y, or
- (ii) (a) W = G or C,
X = A or T/U,
Y = A or T/U,
and X ≠ Y, and
(b) a base not selected in (ii)(a) can be inserted into each
sequence at one or more locations, the location of each
insertion being the same in all the sequences;
- (D) up to three bases can be inserted at any location of any of the
sequences or up to three bases can be deleted from any of the
sequences;
- (E) all of the sequences of a said group of oligonucleotides are read 5'
to 3' or are read 3' to 5'; and

wherein each oligonucleotide of a said set has a sequence of at least ten
contiguous bases of the sequence on which it is based, provided that:

- (F) (I) the quotient of the sum of G and C divided by the sum of A, T/U,
G and C for all combined sequences of the set is between about
0.1 and 0.40 and said quotient for each sequence of the set does
not vary from the quotient for the combined sequences by more
than 0.2; and
- (II) for any phantom sequence generated from any pair of first and
second sequences of the set L_1 and L_2 in length, respectively,
by selection from the first and second sequences of identical
bases in identical sequence with each other:
- (i) any consecutive sequence of bases in the phantom
sequence which is identical to a consecutive sequence of
bases in each of the first and second sequences from
which it is generated is less than $((3/4 \times L) - 1)$ bases
in length;
- (ii) the phantom sequence, if greater than or equal to $(5/6 \times$
 $L)$ in length, contains at least three
insertions/deletions or mismatches when compared to the
first and second sequences from which it is generated;

- 63 -

and

(iii) the phantom sequence is not greater than or equal to
 $(11/12 \times L)$ in length;

where $L = L_1$, or if $L_1 \neq L_2$, where L is the greater of L_1 and L_2 ;
and

wherein any base present may be substituted by an analogue thereof.

2. The composition of claim 1, wherein the composition includes at least ten said molecules, or at least eleven said molecules, or at least twelve said molecules, or at least thirteen said molecules, or at 5 least fourteen said molecules, or at least fifteen said molecules, or at least sixteen said molecules, or at least seventeen said molecules, or at least eighteen said molecules, or at least nineteen said molecules, or at least twenty said molecules, or at least twenty-one said molecules, or at least twenty-two said molecules, or at least 10 twenty-three said molecules, or at least twenty-four said molecules, or at least twenty-five said molecules, or at least twenty-six said molecules, or at least twenty-seven said molecules, or at least twenty-eight said molecules, or at least twenty-nine said molecules, or at least 15 thirty said molecules, or at least thirty-one said molecules, or at least thirty-two said molecules, or at least thirty-three said molecules, or at least thirty-four said molecules, or at least thirty-five said molecules, or at least thirty-six said molecules, or at least thirty-seven said molecules, or at least thirty-eight said molecules, or at least thirty-nine said molecules, or at least forty said 20 molecules, or at least forty-one said molecules, or at least forty-two said molecules, or at least forty-three said molecules, or at least forty-four said molecules, or at least forty-five said molecules, or at least forty-six said molecules, or at least forty-seven said molecules, or at least forty-eight said molecules, or at least forty-nine said 25 molecules, or at least fifty said molecules, or at least sixty said molecules, or at least seventy said molecules, or at least eighty said molecules, or at least ninety said molecules, or at least one hundred said molecules, or at least one hundred and ten said molecules, or at least one hundred and twenty said molecules, or at least one hundred and one hundred and thirty said molecules, or at least one hundred and forty said molecules, or at least one hundred and fifty said molecules, or at least 30 one hundred and sixty said molecules, or at least one hundred and seventy said molecules, or at least one hundred and eighty said molecules,

- 64 -

molecules, or at least one hundred and ninety said molecules, or at least two hundred said molecules.

3. The composition of claim 1, wherein said set of oligonucleotides is
5 based on the sequences tested in Example 2, as set out in Table IA.

4. The composition of claim 3, wherein the composition includes at least ten said molecules, or at least eleven said molecules, or at least twelve said molecules, or at least thirteen said molecules, or at
10 least fourteen said molecules, or at least fifteen said molecules, or at least sixteen said molecules, or at least seventeen said molecules, or at least eighteen said molecules, or at least nineteen said
15 molecules, or at least twenty said molecules, or at least twenty-one said molecules, or at least twenty-two said molecules, or at least twenty-three said molecules, or at least twenty-four said molecules, or at least twenty-five said molecules, or at least twenty-six said
20 molecules, or at least twenty-seven said molecules, or at least twenty-eight said molecules, or at least twenty-nine said molecules, or at least thirty said molecules, or at least thirty-one said molecules, or at least thirty-two said molecules, or at least thirty-three said
25 molecules, or at least thirty-four said molecules, or at least thirty-five said molecules, or at least thirty-six said molecules, or at least thirty-seven said molecules, or at least thirty-eight said molecules, or at least thirty-nine said molecules, or at least forty said
30 molecules, or at least forty-one said molecules, or at least forty-two said molecules, or at least forty-three said molecules, or at least forty-four said molecules, or at least forty-five said molecules, or at least forty-six said molecules, or at least forty-seven said molecules, or at least forty-eight said molecules, or at least forty-nine said
35 molecules, or at least fifty said molecules, or at least sixty said molecules, or at least seventy said molecules, or at least eighty said molecules, or at least ninety said molecules, or at least one hundred said molecules.

5. The composition of claim 1 or claim 2, wherein:

(G) for the group of 24mer sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each

20 = TATA, each 21 = TAAT and each 22 = ATAT, for the group of sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, under a defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 30% of the degree of hybridization between said sequence and its complement, for all oligonucleotides of the set, the maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 50% of the degree of hybridization of the oligonucleotide and its complement.

6. The composition of claim 3 or claim 4, wherein:

- (G) for the group of 24mer sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, for the group of sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, under a defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 30% of the degree of hybridization between said sequence and its complement, for all oligonucleotides of the set, the maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 50% of the degree of hybridization of the oligonucleotide and its complement.

7. The composition of claim 5 wherein, in (G) under said defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence does not exceed 30% of the degree of hybridization between said sequence and its

- 66 -

complement, the degree of hybridization between each sequence and its complement varies by a factor of between 1 and 10, more preferably between 1 and 9, and more preferably between 1 and 8.

5 8. The composition of claim 6 wherein, in (G) under said defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence does not exceed 30% of the degree of hybridization between said sequence and its complement, the degree of hybridization between each sequence and its
10 complement varies by a factor of between 1 and 10, more preferably between 1 and 9, and more preferably between 1 and 8.

9. The composition of claim 7 wherein the maximum degree of hybridization in (G) between a sequence and any complement of a
15 different sequence does not exceed 25%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 20%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 15%, more
20 preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 11%.

10. The composition of claim 8 wherein the maximum degree of hybridization in (G) between a sequence and any complement of a
25 different sequence does not exceed 25%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 20%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 15%, more
30 preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 11%.

35 11. The composition of claim 7 wherein under said defined set of conditions of (G), the maximum degree of hybridization between a sequence and a complement of any other sequence of the set is no more than 15% greater than the maximum degree of hybridization between a sequence and any complement of a different sequence of the said group

- 67 -

of 24mer sequences, more preferably no more than 10% greater, more preferably no more than 5% greater.

12. The composition of claim 8 wherein under said defined set of
5 conditions of (G), the maximum degree of hybridization between a
sequence and a complement of any other sequence of the set is no more
than 15% greater than the maximum degree of hybridization between a
sequence and any complement of a different sequence of the said group
of 24mer sequences, more preferably no more than 10% greater, more
10 preferably no more than 5% greater.

13. The composition of claim 9 wherein under said defined set of
conditions of (G), the maximum degree of hybridization between a
sequence and a complement of any other sequence of the set is no more
15 than 15% greater than the maximum degree of hybridization between a
sequence and any complement of a different sequence of the said group
of 24mer sequences, more preferably no more than 10% greater, more
preferably no more than 5% greater.

20 14. The composition of claim 10 wherein under said defined set of
conditions of (G), the maximum degree of hybridization between a
sequence and a complement of any other sequence of the set is no more
than 15% greater than the maximum degree of hybridization between a
sequence and any complement of a different sequence of the said group
25 of 24mer sequences, more preferably no more than 10% greater, more
preferably no more than 5% greater.

15. The composition of any of claims 5, 7, 9, 11 or 13, wherein said
defined set of conditions results in a level of hybridization that is
30 the same as the level of hybridization obtained when hybridization
conditions include 0.2 M NaCl, 0.1 M Tris, 0.08% Triton X-100, pH 8.0
at 37°C.

35 16. The composition of any of claims 6, 8, 10, 12 or 14, wherein said
defined set of conditions results in a level of hybridization that is
the same as the level of hybridization obtained when hybridization
conditions include 0.2 M NaCl, 0.1 M Tris, 0.08% Triton X-100, pH 8.0
at 37°C.

- 68 -

17. The composition of claim 15 wherein, in (G) said defined set of conditions includes the group of 24mer sequences of (G) being covalently linked to beads.

5 18. The composition of claim 16 wherein, in (G) said defined set of conditions includes the group of 24mer sequences of (G) being covalently linked to beads.

10 19. The composition of claim 17 or 18 wherein, in (G) for the group of 24mers the maximum degree of hybridization between a sequence and any complement of a different sequence does not exceed 15% of the degree of hybridization between said sequence and its complement and the degree of hybridization between each sequence and its complement varies by a factor of between 1 and 9; and for all oligonucleotides of the set, the 15 maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 20% of the degree of hybridization of the oligonucleotide and its complement.

20 20. The composition of any of claims 1 to 19, wherein each of the 4mers represented by numerals 1 to 22 is selected from the group of 4mers consisting of WXXX, WXXY, WXYX, WXYY, WYXX, WYXY, WYYX, WYYY, XWXX, XWXY, XWYY, XXWX, XXWY, XXXW, XXXY, XYWX, XYWW, XYYW, XYXX, YWXY, YWYX, YWYY, YXWX, YXWY, YXXW, YXYW, YYWX, YYWW, YYXW, and 25 YYYW.

21. The composition of claim 20, wherein each of the 4mers represented by numeral 1 are identical to each other, each of the 4mers represented by numeral 2 are identical to each other, each of the 4mers represented by numeral 3 are identical to each other, each of the 4mers represented by numeral 4 are identical to each other, each of the 4mers represented by numeral 5 are identical to each other, each of the 4mers represented by numeral 6 are identical to each other, each of the 4mers represented by numeral 7 are identical to each other, each of the 4mers represented by numeral 8 are identical to each other, each of the 4mers represented by numeral 9 are identical to each other, each of the 4mers represented by numeral 10 are identical to each other, each of the 4mers represented by numeral 11 are identical to each other, each of the 4mers represented by numeral 12 are identical to each other, each of the 4mers represented by numeral 13 are identical to each other, each 30 35 40

of the 4mers represented by numeral 14 are identical to each other,
each of the 4mers represented by numeral 15 are identical to each
other, each of the 4mers represented by numeral 16 are identical to
each other, each of the 4mers represented by numeral 17 are identical
5 to each other, each of the 4mers represented by numeral 18 are
identical to each other, each of the 4mers represented by numeral 19
are identical to each other, each of the 4mers represented by numeral
20 are identical to each other, each of the 4mers represented by
numeral 21 are identical to each other, and each of the 4mers
10 represented by numeral 22 are identical to each other.

22. The composition of claim 20, wherein at least one of the 4mers
represented by the numeral 1 has the sequence WXYX, at least one of the
4mers represented by the numeral 2 has the sequence YWXY, at least one
15 of the 4mers represented by the numeral 3 has the sequence XXXW, at
least one of the 4mers represented by the numeral 4 has the sequence
YWYX, at least one of the 4mers represented by the numeral 5 has the
sequence WYXY, at least one of the 4mers represented by the numeral 6
has the sequence YYWX, at least one of the 4mers represented by the
20 numeral 7 has the sequence YWXX, at least one of the 4mers represented
by the numeral 8 has the sequence WYXX, at least one of the 4mers
represented by the numeral 9 has the sequence XYYW, at least one of the
4mers represented by the numeral 10 has the sequence XYWX, at least one
25 of the 4mers represented by the numeral 11 has the sequence YYXW, at
least one of the 4mers represented by the numeral 12 has the sequence
WYYX, at least one of the 4mers represented by the numeral 13 has the
sequence XYXW, at least one of the 4mers represented by the numeral 14
has the sequence WYYY, at least one of the 4mers represented by the
30 numeral 15 has the sequence WXYW, at least one of the 4mers represented
by the numeral 16 has the sequence WYXW, at least one of the 4mers
represented by the numeral 17 has the sequence WXXW, at least one of
the 4mers represented by the numeral 18 has the sequence WYYW, at least
one of the 4mers represented by the numeral 19 has the sequence XYYX,
35 at least one of the 4mers represented by the numeral 20 has the
sequence YXYX, at least one of the 4mers represented by the numeral 21
has the sequence YXXX, and at least one of the 4mers represented by the
numeral 22 has the sequence XYXY.

23. The composition of claim 22, wherein in each 1 = WXYX, each 2 =
40 YWXY, each 3 = XXXW, each 4 = YWYX, each 5 = WYXY, each 6 = YYWX, each

- 70 -

7 = YWXX, each 8 = WYXX, each 9 = XYYW, each 10 = XYWX, each 11 = YYXW,
each 12 = WYYX, each 13 = XYXW, each 14 = WYYY, each 15 = WXYW, each 16
= WYXW, each 17 = WXXW, each 18 = WYYW, each 19 = XYYX, each 20 = YXYX,
each 21 = YXXY and each 22 = XYXY.

5

24. The composition of any of claims 1, wherein a said group of sequences is based on the sequences having sequence identifiers 1 to 173 as set out in Table IA, and wherein each of the 4mers represented by numerals 1 to 14 in (A) is selected from the group of 4mers
10 consisting of WXYY, YWXY, XXXW, YWYX, WYXY, YYWX, YWXX, WYXX, XYYW,
XYWX, YYXW, WYYX, XYXW, and WYYY.

25. The composition of claim 24, wherein the composition includes at least ten said molecules, or at least eleven said molecules, or at least twelve said molecules, or at least thirteen said molecules, or at least fourteen said molecules, or at least fifteen said molecules, or at least sixteen said molecules, or at least seventeen said molecules, or at least eighteen said molecules, or at least nineteen said molecules, or at least twenty said molecules, or at least twenty-one said molecules, or at least twenty-two said molecules, or at least twenty-three said molecules, or at least twenty-four said molecules, or at least twenty-five said molecules, or at least twenty-six said molecules, or at least twenty-seven said molecules, or at least twenty-eight said molecules, or at least twenty-nine said molecules, or at least thirty said molecules, or at least thirty-one said molecules, or at least thirty-two said molecules, or at least thirty-three said molecules, or at least thirty-four said molecules, or at least thirty-five said molecules, or at least thirty-six said molecules, or at least thirty-seven said molecules, or at least thirty-eight said molecules,
30 or at least thirty-nine said molecules, or at least forty said molecules, or at least forty-one said molecules, or at least forty-two said molecules, or at least forty-three said molecules, or at least forty-four said molecules, or at least forty-five said molecules, or at least forty-six said molecules, or at least forty-seven said molecules,
35 or at least forty-eight said molecules, or at least forty-nine said molecules, or at least fifty said molecules, or at least sixty said molecules, or at least seventy said molecules, or at least eighty said molecules, or at least ninety said molecules, or at least one hundred said molecules, or at least one hundred and ten said molecules, or at least one hundred and twenty said molecules, or at least one hundred

- 71 -

and thirty said molecules, or at least one hundred and forty said molecules, or at least one hundred and fifty said molecules, or at least one hundred and sixty said molecules, or at least one hundred and seventy said molecules, or at least one hundred and eighty said
5 molecules, or at least one hundred and ninety said molecules, or at least two hundred said molecules.

26. The composition of claim 24 or claim 25, wherein:

(G) for the group of 24mer sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, under a defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 30% of the degree of hybridization between said sequence and its complement, for all oligonucleotides of the set, the maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 50% of the degree of hybridization of the oligonucleotide and its complement.

27. The composition of claim 26 wherein, in (G) under said defined set
10 of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence does not exceed 30% of the degree of hybridization between said sequence and its complement, the degree of hybridization between each sequence and its complement varies by a factor of between 1 and 10, more preferably
15 between 1 and 9, and more preferably between 1 and 8.

28. The composition of claim 27 wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 25%, more preferably wherein the
20 maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 20%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 15%, more preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 11%.
25

- 72 -

29. The composition of claim 27 or claim 28 wherein under said defined set of conditions of (G), the maximum degree of hybridization between a sequence and a complement of any other sequence of the set is no more than 15% greater than the maximum degree of hybridization between a sequence and any complement of a different sequence of the said group of 24mer sequences, more preferably no more than 10% greater, more preferably no more than 5% greater.

10 30. The composition of any of claims 26 to 29, wherein said defined set of conditions results in a level of hybridization that is the same as the level of hybridization obtained when hybridization conditions include 0.2 M NaCl, 0.1 M Tris, 0.08% Triton X-100, pH 8.0 at 37°C.

15 31. The composition of claim 30 wherein, in (G) said defined set of conditions includes the group of 24mer sequences of (G) being covalently linked to beads.

20 32. The composition of any of claims 24 to 31, wherein each of the 4mers represented by numeral 1 are identical to each other, each of the 4mers represented by numeral 2 are identical to each other, each of the 4mers represented by numeral 3 are identical to each other, each of the 4mers represented by numeral 4 are identical to each other, each of the 4mers represented by numeral 5 are identical to each other, each of the 4mers represented by numeral 6 are identical to each other, each of the 4mers represented by numeral 7 are identical to each other, each of the 4mers represented by numeral 8 are identical to each other, each of the 4mers represented by numeral 9 are identical to each other, each of the 4mers represented by numeral 10 are identical to each other, each of the 4mers represented by numeral 11 are identical to each other, each of the 4mers represented by numeral 12 are identical to each other, each of the 4mers represented by numeral 13 are identical to each other, and each of the 4mers represented by numeral 14 are identical to each other.

35 33. The composition of claim 24 to 31, wherein at least one of the 4mers represented by the numeral 1 has the sequence WXYY, at least one of the 4mers represented by the numeral 2 has the sequence YWXY, at least one of the 4mers represented by the numeral 3 has the sequence XXXW, at least one of the 4mers represented by the numeral 4 has the sequence YWYX, at least one of the 4mers represented by the numeral 5

has the sequence WYXY, at least one of the 4mers represented by the numeral 6 has the sequence YYWX, at least one of the 4mers represented by the numeral 7 has the sequence YWXX, at least one of the 4mers represented by the numeral 8 has the sequence WYXX, at least one of the 5 4mers represented by the numeral 9 has the sequence XYYW, at least one of the 4mers represented by the numeral 10 has the sequence XYWX, at least one of the 4mers represented by the numeral 11 has the sequence YYXW, at least one of the 4mers represented by the numeral 12 has the sequence WYYX, at least one of the 4mers represented by the numeral 13 10 has the sequence XYXW, and at least one of the 4mers represented by the numeral 14 has the sequence WYYY.

34. The composition of claim 33, wherein each 1 = WXYY, each 2 = YWXY, each 3 = XXXW, each 4 = YWYX; each 5 = WYXY, each 6 = YYWX, each 7 = 15 YYXW, each 8 = WYXX, each 9 = XYYW, each 10 = XYWX, each 11 = YYXW, each 12 = WYYX, each 13 = XYXW, and each 14 = WYYY.

35. The composition of claim 1, wherein a said group of sequences is based on those sequences having sequence identifiers 1 to 100 as set 20 out in Table IA and wherein each of the 4mers represented by numerals 1 to 10 in (A) is selected from the group of 4mers consisting of WXYY, YWXY, XXXW, YWYX, WYXY, YYWX, YWXX, WYXX, XYYW, and XYWX.

36. The composition of claim 35, wherein the composition includes at 25 least ten said molecules, or at least eleven said molecules, or at least twelve said molecules, or at least thirteen said molecules, or at least fourteen said molecules, or at least fifteen said molecules, or at least sixteen said molecules, or at least seventeen said molecules, or at least eighteen said molecules, or at least nineteen said 30 molecules, or at least twenty said molecules, or at least twenty-one said molecules, or at least twenty-two said molecules, or at least twenty-three said molecules, or at least twenty-four said molecules, or at least twenty-five said molecules, or at least twenty-six said molecules, or at least twenty-seven said molecules, or at least twenty-eight said molecules, or at least twenty-nine said molecules, or at 35 least thirty said molecules, or at least thirty-one said molecules, or at least thirty-two said molecules, or at least thirty-three said molecules, or at least thirty-four said molecules, or at least thirty-five said molecules, or at least thirty-six said molecules, or at least thirty-seven said molecules, or at least thirty-eight said molecules,

- 74 -

or at least thirty-nine said molecules, or at least forty said
molecules, or at least forty-one said molecules, or at least forty-two
said molecules, or at least forty-three said molecules, or at least
forty-four said molecules, or at least forty-five said molecules, or at
5 least forty-six said molecules, or at least forty-seven said molecules,
or at least forty-eight said molecules, or at least forty-nine said
molecules, or at least fifty said molecules, or at least sixty said
molecules, or at least seventy said molecules, or at least eighty said
molecules, or at least ninety said molecules, or at least one hundred
10 said molecules.

37. The composition of claim 35 or claim 36, wherein:

(G) for the group of 24mer sequences in which each 1 = GATT, each 2 =
TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA,
each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 =
TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, under a defined
set of conditions in which the maximum degree of hybridization between
a sequence and any complement of a different sequence of the group of
24mer sequences does not exceed 30% of the degree of hybridization
between said sequence and its complement, for all oligonucleotides of
the set, the maximum degree of hybridization between an
oligonucleotide and a complement of any other oligonucleotide of the
set does not exceed 50% of the degree of hybridization of the
oligonucleotide and its complement.

38. The composition of claim 37 wherein, in (G) under said defined set
of conditions in which the maximum degree of hybridization between a
15 sequence and any complement of a different sequence does not exceed 30%
of the degree of hybridization between said sequence and its
complement, the degree of hybridization between each sequence and its
complement varies by a factor of between 1 and 10, more preferably
between 1 and 9, and more preferably between 1 and 8.

20

39. The composition of claim 38 wherein the maximum degree of
hybridization in (G) between a sequence and any complement of a
different sequence does not exceed 25%, more preferably wherein the
maximum degree of hybridization in (G) between a sequence and any
25 complement of a different sequence does not exceed 20%, more preferably
wherein the maximum degree of hybridization in (G) between a sequence
and any complement of a different sequence does not exceed 15%, more

- 75 -

preferably wherein the maximum degree of hybridization in (G) between a sequence and any complement of a different sequence does not exceed 11%.

5 40. The composition of claim 38 or claim 39 wherein under said defined set of conditions of (G), the maximum degree of hybridization between a sequence and a complement of any other sequence of the set is no more than 15% greater than the maximum degree of hybridization between a sequence and any complement of a different sequence of the said group
10 of 24mer sequences, more preferably no more than 10% greater, more preferably no more than 5% greater.

15 41. The composition of any of claims 40 to 37, wherein said defined set of conditions results in a level of hybridization that is the same as the level of hybridization obtained when hybridization conditions include 0.2 M NaCl, 0.1 M Tris, 0.08% Triton X-100, pH 8.0 at 37°C.

20 42. The composition of claim 41 wherein, in (G) said defined set of conditions includes the group of 24mer sequences of (G) being covalently linked to beads.

25 43. The composition of any of claims 34 to 41, wherein each of the 4mers represented by numeral 1 are identical to each other, each of the 4mers represented by numeral 2 are identical to each other, each of the 4mers represented by numeral 3 are identical to each other, each of the 4mers represented by numeral 4 are identical to each other, each of the 4mers represented by numeral 5 are identical to each other, each of the 4mers represented by numeral 6 are identical to each other, each of the 4mers represented by numeral 7 are identical to each other, each of the 4mers represented by numeral 8 are identical to each other, each of the 4mers represented by numeral 9 are identical to each other, and each of the 4mers represented by numeral 10 are identical to each other.

35 44. The composition of claim 43, wherein at least one of the 4mers represented by the numeral 1 has the sequence WXYY, at least one of the 4mers represented by the numeral 2 has the sequence YWXY, at least one of the 4mers represented by the numeral 3 has the sequence XXXW, at least one of the 4mers represented by the numeral 4 has the sequence YWYX, at least one of the 4mers represented by the numeral 5 has the sequence WYXY, at least one of the 4mers represented by the numeral 6

- 76 -

has the sequence YYWX, at least one of the 4mers represented by the numeral 7 has the sequence YWXX, at least one of the 4mers represented by the numeral 8 has the sequence WYXX, at least one of the 4mers represented by the numeral 9 has the sequence XYYW, and at least one of the 4mers represented by the numeral 10 has the sequence XYWX.

5

45. The composition of claim 44, wherein each 1 = WXYY, each 2 = YWXY, each 3 = XXXW, each 4 = YWYX, each 5 = WYXY, each 6 = YYWX, each 7 = YWXX, each 8 = WYXX, each 9 = XYYW, and each 10 = XYWX.

10

46. The composition of any preceding claim, wherein in (C)(i)(a): W = one of G and C; X = one of A and T/U; and Y = one of A and T/U.

15

47. The composition of claim 46, wherein in (C)(i)(a): W = G; X = one of A, and T/U; and Y = one of A and T/U.

48. The composition of any preceding claim, wherein W = G; X = A; and Y = T/U.

20

49. The composition of any preceding claim, wherein in (F)(I), said quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.1.

25

50. The composition of claim 51, wherein in (F)(I) , said quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.05.

30

51. The composition of claim 50, wherein in (F)(I) , said quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.01.

35

52. The composition of any preceding claim, wherein in (F)(I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.15 and 0.35.

53. The composition of claim 52, wherein in (F)(I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.2 and 0.3.

- 77 -

54. The composition of claim 53, wherein in (F)(I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.21 and 0.29.

5 55. The composition of claim 54, wherein in (F)(I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.22 and 0.28.

10 56. The composition of claim 55, wherein in (F)(I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.23 and 0.27.

15 57. The composition of claim 56, wherein in (F)(I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.24 and 0.26.

58. The composition of claim 57, wherein in (F)(I) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is 0.25.

20 59. The composition of any preceding claim, wherein in (D) up to two bases can be inserted at any location of any of the sequences or up to two bases can be deleted from any of the sequences.

25 60. The composition of claim 59, wherein in (D) one base can be inserted at any location of any of the sequences or one base can be deleted from any of the sequences.

30 61. The composition of claim 60, wherein in (D) no base can be inserted at any location of any of the sequences.

62. The composition of claim 60, wherein in (D) no base can be deleted from any of the sequences.

35 63. The composition of claim 60, wherein in (D) no base can be inserted at or deleted from any location of any of the sequences.

40 64. The composition of any preceding claim, wherein each of the oligonucleotides of a said set has a sequence at least eleven contiguous bases of the sequence on which it is based; or wherein each

of the oligonucleotides of a said set has a sequence at least twelve contiguous bases of the sequence on which it is based; or wherein each of the oligonucleotides of a said set has a sequence at least thirteen contiguous bases of the sequence on which it is based; or wherein each 5 of the oligonucleotides of a said set has a sequence at least fourteen contiguous bases of the sequence on which it is based; or wherein each of the oligonucleotides of a said set has a sequence at least fifteen contiguous bases of the sequence on which it is based; or wherein each 10 of the oligonucleotides of a said set has a sequence at least sixteen contiguous bases of the sequence on which it is based; or wherein each of the oligonucleotides of a said set has a sequence at least seventeen contiguous bases of the sequence on which it is based; or wherein each 15 of the oligonucleotides of a said set has a sequence at least eighteen contiguous bases of the sequence on which it is based; or wherein each of the oligonucleotides of a said set has a sequence at least nineteen contiguous bases of the sequence on which it is based; or wherein each 20 of the oligonucleotides of a said set has a sequence at least twenty contiguous bases of the sequence on which it is based; or wherein each of the oligonucleotides of a said set has a sequence at least twenty-one contiguous bases of the sequence on which it is based; or wherein each 25 of the oligonucleotides of a said set has a sequence at least twenty-two contiguous bases of the sequence on which it is based; or wherein each of the oligonucleotides of a said set has a sequence at least twenty-three contiguous bases of the sequence on which it is based; or wherein each of the oligonucleotides of a said set has a sequence at least twenty-four contiguous bases of the sequence on which it is based.

65. The composition according to any preceding claim, wherein in each 30 oligonucleotide of the set, there is a maximum of six bases other than G between every neighboring pair of G's.

66. The composition according to claim 65, wherein each initial G of an oligonucleotide of the set sequence occupies a position in the 35 terminal selected from a first, second, third, fourth, fifth, sixth or seventh position thereof.

67. The composition according to any preceding claim wherein the 40 contiguous bases of each oligonucleotide of a said set are selected such that the position of the first base of each said oligonucleotide

- 79 -

within the sequence on which it is based is the same for all nucleotides of the set.

68. The composition of any preceding claim, wherein each of the
5 oligonucleotides of a said set is up to thirty bases in length; or each
of the oligonucleotides of a said set is up to twenty-nine bases in
length; or each of the oligonucleotides of a said set is up to twenty-
eight bases in length; or each of the oligonucleotides of a said set is
up to twenty-seven bases in length; or each of the oligonucleotides of a said set is
10 up to twenty-six bases in length; or each of the
oligonucleotides of a said set is up to twenty-five bases in length; or
each of the oligonucleotides of a said set is up to twenty-four bases
in length.
- 15 69. The composition of any preceding claim, wherein each of the
oligonucleotides of a said set has a length of within five bases of the
average length of all of the oligonucleotides in the set; or each of
the oligonucleotides of a said set has a length of within four bases of
the average length of all of the oligonucleotides in the set; or each
20 of the oligonucleotides of a said set has a length of within three
bases of the average length of all of the oligonucleotides in the set;
or each of the oligonucleotides of a said set has a length of within
two bases of the average length of all of the oligonucleotides in the
set; or each of the oligonucleotides of a said set has a length of
25 within one base of the average length of all of the oligonucleotides in
the set.
70. The composition of any preceding claim, wherein in (II)(i), any
consecutive sequence of bases in the phantom sequence which is
30 identical to a consecutive sequence of bases in each of the first and
second sequences from which it is generated is no more $((2/3 \times L) - 1)$
bases in length.
71. The composition of any preceding claim, wherein in (II)(ii), the
35 phantom sequence, if greater than or equal to $(3/4 \times L)$ in length,
contains at least 3 insertions/deletions or mismatches when compared to
the first and second sequences from which it is generated.
72. The composition of claim 71, wherein in (II)(ii), the phantom
40 sequence, if greater than or equal to $(2/3 \times L)$ in length, contains at

- 80 -

least 3 insertions/deletions or mismatches when compared to the first and second sequences from which it is generated.

73. The composition of any preceding claim, wherein in (II)(iii), the 5 phantom sequence is not greater than or equal to $(5/6 \times L)$ in length.

74. The composition of claim 73, wherein in (II)(iii), the phantom sequence is not greater than or equal to $(3/4 \times L)$ in length.

75. A composition comprising molecules for use as tags or tag complements wherein each molecule comprises an oligonucleotide selected from a set of oligonucleotides based on a following group of sequences having the one hundred sequence identifiers of the sequences tested in Example 2 as set out in Table IA.

wherein:

- (A) wherein 1 = WXYZ, each 2 = YWXY, each 3 = XXXW, each 4 = YWYX, each 5 = WYXY, each 6 = YYWX, each 7 = YWXX, each 8 = WYXX, each 9 = XYYW, each 10 = XYWX, each 11 = YYXW, each 12 = WYYX, each 13 = XYXW, each 14 = WYYY, each 15 = WXYW, each 16 = WYXW, each 17 = WXXW, each 18 = WYYW, each 19 = XYYX, each 20 = YXYX, each 21 = YXXY and each 22 = XYXY;
- (B) each of W, X and Y is a base in which either:
 - (i) (a) W = one of A, T/U, G, and C,
X = one of A, T/U, G, and C,
Y = one of A, T/U, G, and C,
and each of W, X and Y is selected so as to be different from all of the others of W, X and Y,
 - (b) an unselected said base of (i)(a) can be substituted any number of times for any one of W, X and Y, or
- (ii) (a) W = G or C,
X = A or T/U,
Y = A or T/U,
and X ≠ Y, and
- (b) a base not selected in (ii)(a) can be inserted into each sequence at one or more locations, the location of each insertion being the same in all the sequences;
- (C) up to three bases can be inserted at any location of any of the sequences or up to three bases can be deleted from any of the sequences;

- 81 -

- (D) all of the sequences of a said group of oligonucleotides are read 5' to 3' or are read 3' to 5'; and

wherein each oligonucleotide of a said set has a sequence of at least ten contiguous bases of the sequence on which it is based, provided that:

- (E) the quotient of the sum of G and C divided by the sum of A, T/U, G and C for all combined sequences of the set is between about 0.1 and 0.40 and said quotient for each sequence of the set does not vary from the quotient for the combined sequences by more than 0.2; and

- (F) for the group of 24mer sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, for the group of sequences in which each 1 = GATT, each 2 = TGAT, each 3 = AAAG, each 4 = TGTA, each 5 = GTAT, each 6 = TTGA, each 7 = TGAA, each 8 = GTAA, each 9 = ATTG, each 10 = ATGA, each 11 = TTAG, each 12 = GTTA, each 13 = ATAG, each 14 = GTTT, each 15 = GATG, each 16 = GTAG, each 17 = GAAG, each 18 = GTTG, each 19 = ATTA, each 20 = TATA, each 21 = TAAT and each 22 = ATAT, under a defined set of conditions in which the maximum degree of hybridization between a sequence and any complement of a different sequence of the group of 24mer sequences does not exceed 30% of the degree of hybridization between said sequence and its complement, for all oligonucleotides of the set, the maximum degree of hybridization between an oligonucleotide and a complement of any other oligonucleotide of the set does not exceed 50% of the degree of hybridization of the oligonucleotide and its complement; wherein any base present may be substituted by an analogue thereof.

76. The composition of claim 75 wherein the contiguous bases of each oligonucleotide of a said set are selected such that the position of the first base of each said oligonucleotide within the sequence on which it is based is the same for all nucleotides of the set.

77. The composition of claim 75 or 76 wherein, subject to the provisos of (E) and (F), each oligonucleotide of a said set comprises a said sequence of twenty-four contiguous bases of the sequence on which it is based.

- 82 -

78. The composition of claim 75 or 76 wherein, subject to the proviso of (F) each oligonucleotide of a said set comprises a said sequence of twenty-four contiguous bases of the sequence on which it is based.

5 79. The composition of any of claims 75 to 81, wherein in (B): W = one of G and C; X = one of A and T/U; and Y = one of A and T/U.

80. The composition of claim 79, wherein in (B): W = G; X = one of A, and T/U; and Y = one of A and T/U.

10 81. The composition of any of claims 75 to 80, wherein the composition includes at least ten said molecules, or at least eleven said molecules, or at least twelve said molecules, or at least thirteen said molecules, or at least fourteen said molecules, or at least fifteen said molecules, or at least sixteen said molecules, or at least seventeen said molecules, or at least eighteen said molecules, or at least nineteen said molecules, or at least twenty said molecules, or at least twenty-one said molecules, or at least twenty-two said molecules, or at least twenty-three said molecules, or at least twenty-four said molecules, or at least twenty-five said molecules, or at least twenty-six said molecules, or at least twenty-seven said molecules, or at least twenty-eight said molecules, or at least twenty-nine said molecules, or at least thirty said molecules, or at least thirty-one said molecules, or at least thirty-two said molecules, or at least thirty-three said molecules, or at least thirty-four said molecules, or at least thirty-five said molecules, or at least thirty-six said molecules, or at least thirty-seven said molecules, or at least thirty-eight said molecules, or at least thirty-nine said molecules, or at least forty said molecules, or at least forty-one said molecules, or at least forty-two said molecules, or at least forty-three said molecules, or at least forty-four said molecules, or at least forty-five said molecules, or at least forty-six said molecules, or at least forty-seven said molecules, or at least forty-eight said molecules, or at least forty-nine said molecules, or at least fifty said molecules, or at least sixty said molecules, or at least seventy said molecules, or at least eighty said molecules, or at least ninety said molecules, or at least one hundred said molecules.

- 83 -

82. A composition of any preceding claim, wherein each molecule is linked to a solid phase support so as to be distinguishable from a mixture of said molecules by hybridization to its complement.

5 83. The composition of claim 82, wherein each molecule is linked to a defined location on a said solid phase support, the defined location for each said molecule being different than the defined location for different other said molecules.

10 84. The composition of claim 82, wherein each said solid phase support is a microparticle and each said molecule is covalently to a different microparticle than each other different said molecule.

15 85. A composition according to any of claims 1 to 84, wherein each said molecule comprises a tag complement.

86. A kit for sorting and identifying polynucleotides, the kit comprising one or more solid phase supports each having one or more spatially discrete regions, each such region having a uniform population of substantially identical tag complements covalently attached, and the tag complements each being selected from the set of oligonucleotides as defined in any of claims 1 to 85.

25 87. A kit according to claim 86, wherein there is a tag complement for each said oligonucleotide of a said composition.

30 88. A kit according to claim 86 or 87 wherein said one or more solid phase supports is a planar substrate and wherein said one or more spatially discrete regions is a plurality of spatially addressable regions.

89. A kit according to any of claims 86 to 88 wherein said one or more solid phase supports is a plurality of microparticles.

35 90. A kit according to claim 89 wherein said microparticles each have a diameter in the range of from 5 to 40 μm .

40 91. A kit according to claim 89 or 90, wherein each microparticle is spectrophotometrically unique from each other microparticle having a different oligonucleotide attached thereto.

- 84 -

92. A method of analyzing a biological sample comprising a biological sequence for the presence of a mutation or polymorphism at a locus of the nucleic acid, the method comprising:

- (A) amplifying the nucleic acid molecule in the presence of a first primer having a 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements as defined in claim 85 to form an amplified molecule with a 5'-end with a sequence complementary to the sequence of the tag;
- (B) extending the amplified molecule in the presence of a polymerase and a second primer having 5'-end complementary the 3'-end of the amplified sequence, with the 3'-end of the second primer extending to immediately adjacent said locus, in the presence of a plurality of nucleoside triphosphate derivatives each of which is: (i) capable of incorporation during transcription by the polymerase onto the 3'-end of a growing nucleotide strand; (ii) causes termination of polymerization; and (iii) capable of differential detection, one from the other, wherein there is a said derivative complementary to each possible nucleotide present at said locus of the amplified sequence;
- (C) specifically hybridizing the second primer to a tag complement having the tag complement sequence of (A); and
- (D) detecting the nucleotide derivative incorporated into the second primer in (B) so as to identify the base located at the locus of the nucleic acid.

5

93. A method of analyzing a biological sample comprising a plurality of nucleic acid molecules for the presence of a mutation or polymorphism at a locus of each nucleic acid molecule, for each nucleic acid molecule, the method comprising:

- (A) amplifying the nucleic acid molecule in the presence of a first primer having a 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements as defined in claim 85 to form an amplified molecule with a 5'-end with a sequence complementary to the sequence of the tag;
- (B) extending the amplified molecule in the presence of a polymerase and a second primer having 5'-end complementary the 3'-end of the amplified sequence, the 3'-end of the second primer extending to immediately adjacent said locus, in the presence of a plurality of nucleoside triphosphate derivatives each of which is: (i) capable of incorporation during transcription by the polymerase onto the 3'-end of

a growing nucleotide strand; (ii) causes termination of polymerization; and (iii) capable of differential detection, one from the other, wherein there is a said derivative complementary to each possible nucleotide present at said locus of the amplified molecule;

- (C) specifically hybridizing the second primer to a tag complement having the tag complement sequence of (A); and
- (D) detecting the nucleotide derivative incorporated into the second primer in (B) so as to identify the base located at the locus of the nucleic acid;

wherein each tag of (A) is unique for each nucleic acid molecule and steps (A) and (B) are carried out with said nucleic molecules in the presence of each other.

94. A method of analyzing a biological sample comprising a plurality of double stranded complementary nucleic acid molecules for the presence of a mutation or polymorphism at a locus of each nucleic acid
5 molecule, for each nucleic acid molecule, the method comprising:

- (A) amplifying the double stranded molecule in the presence of a pair of first primers, each primer having an identical 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements as defined in claim 85 to form amplified molecules with 5'-ends with a sequence complementary to the sequence of the tag;
- (B) extending the amplified molecules in the presence of a polymerase and a pair of second primers each second primer having a 5'-end complementary a 3'-end of the amplified sequence, the 3'-end of each said second primer extending to immediately adjacent said locus, in the presence of a plurality of nucleoside triphosphate derivatives each of which is:
 - (i) capable of incorporation during transcription by the polymerase onto the 3'-end of a growing nucleotide strand; (ii) causes termination of polymerization; and (iii) capable of differential detection, one from the other;
- (C) specifically hybridizing each of the second primers to a tag complement having the tag complement sequence of (A); and
- (D) detecting the nucleotide derivative incorporated into the second primers in (B) so as to identify the base located at said locus; wherein the sequence of each tag of (A) is unique for each nucleic acid molecule and steps (A) and (B) are carried out with said nucleic molecules in the presence of each other.

- 86 -

95. A method of analyzing a biological sample comprising a plurality of nucleic acid molecules for the presence of a mutation or polymorphism at a locus of each nucleic acid molecule, for each nucleic acid molecule, the method comprising:

- (a) hybridizing the molecule and a primer, the primer having a 5'-sequence having the sequence of a tag complementary to the sequence of a tag complement belonging to a family of tag complements as defined in claim 85 and a 3'-end extending to immediately adjacent the locus;
- (b) enzymatically extending the 3'-end of the primer in the presence of a plurality of nucleoside triphosphate derivatives each of which is: (i) capable of enzymatic incorporation onto the 3'-end of a growing nucleotide strand; (ii) causes termination of said extension; and (iii) capable of differential detection, one from the other, wherein there is a said derivative complementary to each possible nucleotide present at said locus;
- (c) specifically hybridizing the extended primer formed in step (b) to a tag complement having the tag complement sequence of (a); and
- (d) detecting the nucleotide derivative incorporated into the primer in step (b) so as to identify the base located at the locus of the nucleic acid molecule;

wherein each tag of (a) is unique for each nucleic acid molecule and steps (a) and (b) are carried out with said nucleic molecules in the presence of each other.

5

96. The method of claim 93 wherein each said derivative is a dideoxy nucleoside triphosphate.

10 97. The method of claim 95, wherein each respective complement is attached as a uniform population of substantially identical complements in a spatially discrete region on one or more said solid phase supports.

15 98. The method of claim 97, each said tag complement comprises a label, each such label being different for respective complements, and step (d) includes detecting the presence of the different labels for respective hybridization complexes of bound tags and tag complements.

20 99. The hybridized molecule and primer of step (A) of any of claims 95 to 98.

- 87 -

100. A method of determining the presence of a target suspected of being contained in a mixture, the method comprising the steps of:
- (i) labelling the target with a first label;
 - (ii) providing a first detection moiety capable of specific binding to the target and including a first tag;
 - (iii) exposing a sample of the mixture to the detection moiety under conditions suitable to permit (or cause) said specific binding of the molecule and target;
 - (iv) providing a family of tag complements as defined in claim 85 wherein the family contains a first tag complement having a sequence complementary to that of the first tag;
 - (v) exposing the sample to the family of tag complements under conditions suitable to permit (or cause) specific hybridization of the first tag and its tag complement;
 - (vi) determining whether a said first detection moiety hybridized to a first said tag complement is bound to a said labelled target in order to determine the presence or absence said target in the mixture.

101. The method of claim 100 wherein said first tag complement is linked to a solid support at a specific location of the support and step (vi) includes detecting the presence the first label at said specified location.

102. The method of claim 100 wherein said first tag complement comprises a second label and step (vi) includes detecting the presence of the first and second labels in a hybridized complex of the moiety and the first tag complement.

103. The method of claim 100 wherein said target is selected from the group consisting of organic molecules, antigens, proteins, polypeptides, antibodies and nucleic acids.

104. The method of claim 103, wherein said target is an antigen and said first molecule is an antibody specific for said antigen.

20

105. The method of claim 104, wherein the antigen is a polypeptide or protein and the labelling step includes conjugation of fluorescent molecules, digoxigenin, biotinylation and the like.

- 88 -

106. The method of claim 105, wherein said target is a nucleic acid and the labelling step includes incorporation of fluorescent molecules, radiolabelled nucleotide, digoxigenin, biotinylation and the like.

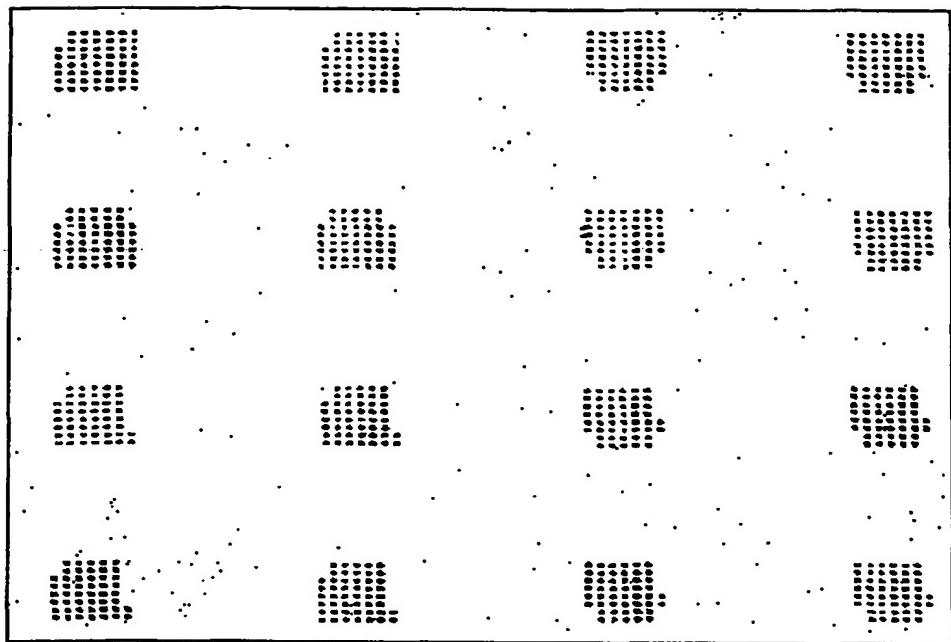


Figure 1b

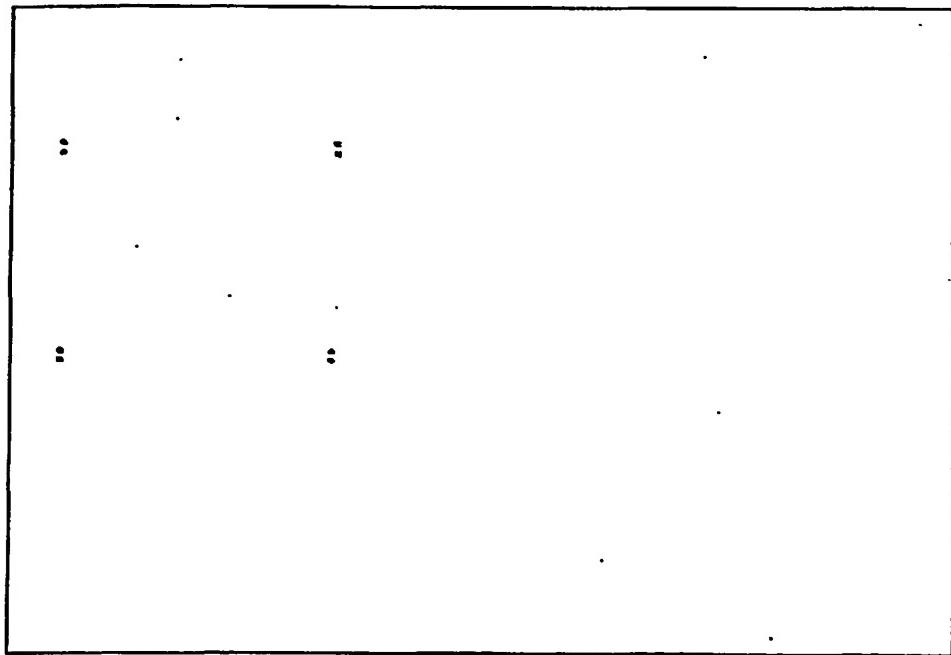
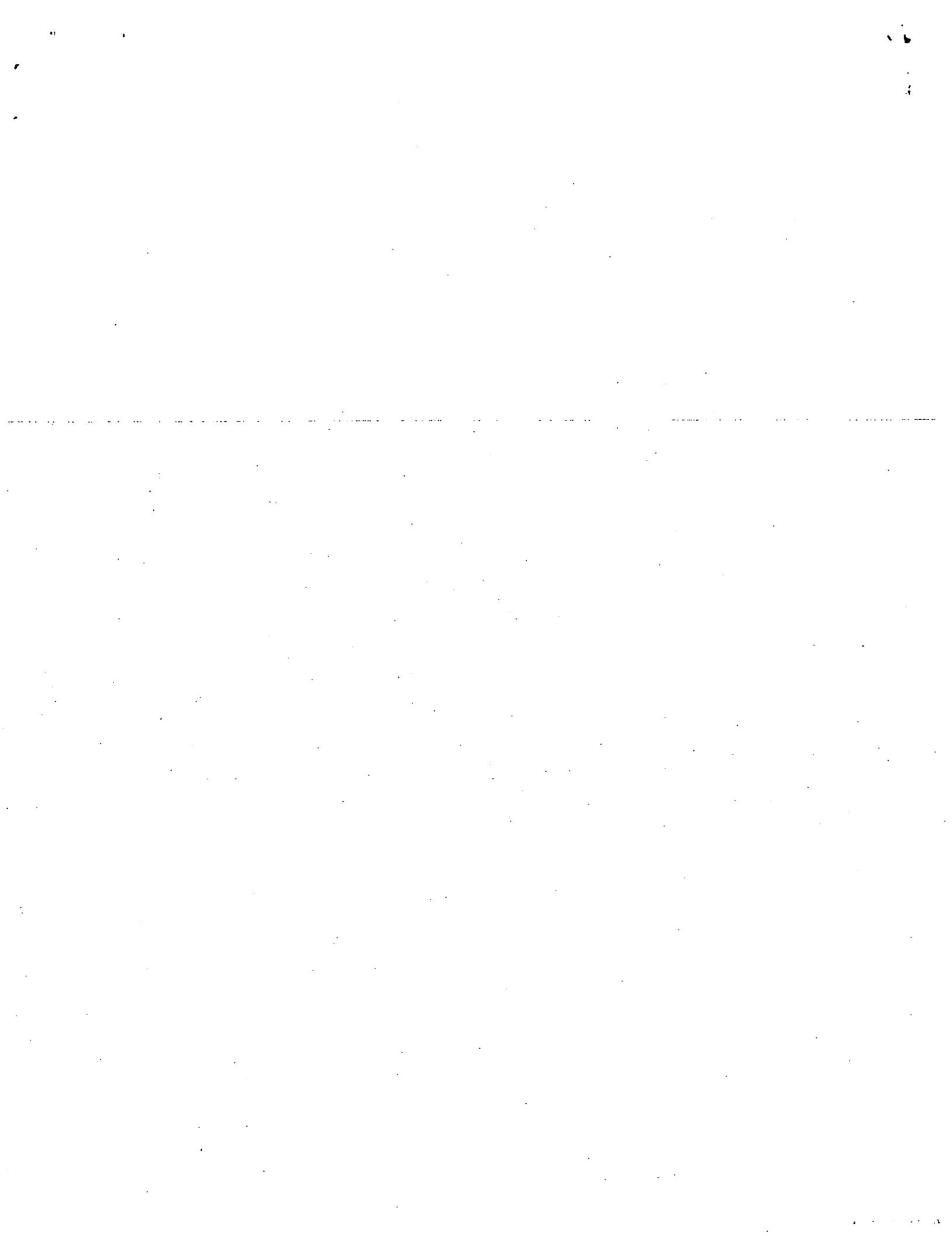


Figure 1a



INTERNATIONAL SEARCH REPORT

Info on patient family members

International Application No.

PCT/CN 12/00087

Patent document cited in search report		Publication date		Patent family member(s)		Publication date
WO 0058516	A	05-10-2000	EP JP WO	1165839 A2 2002539849 T 0058516 A2		02-01-2002 26-11-2002 05-10-2000
EP 0799897	A	08-10-1997	US EP	6458530 B1 0799897 A1		01-10-2002 08-10-1997
US 5654413	A	05-08-1997	US AU AU AU CA CZ EP FI HU JP NO US US US US US US US US US US US US US US US US AU DE DE EP EP WO US US US US	5604097 A 712929 B2 4277896 A 5266399 A 2202167 A1 9700866 A3 0793718 A1 971473 A 77916 A2 10507357 T 971644 A 6352828 B1 6138077 A 6280935 B1 6172218 B1 9612014 A1 6235475 B1 6172214 B1 6150516 A 5635400 A 5846719 A 3946195 A 69513997 D1 69513997 T2 0786014 A1 0952216 A2 9612039 A1 6140489 A 5695934 A 5863722 A		18-02-1997 18-11-1999 06-05-1996 09-12-1999 25-04-1996 17-09-1997 10-09-1997 04-06-1997 28-10-1998 21-07-1998 02-06-1997 05-03-2002 24-10-2000 28-08-2001 09-01-2001 25-04-1996 22-05-2001 09-01-2001 21-11-2000 03-06-1997 08-12-1998 06-05-1996 20-01-2000 27-07-2000 30-07-1997 27-10-1999 25-04-1996 31-10-2000 09-12-1997 26-01-1999
EP 0698792	A	28-02-1996	CA EP JP US WO	2162568 A1 0698792 A1 3197277 B2 5789165 A 9427150 A1		24-11-1994 28-02-1996 13-08-2001 04-08-1998 24-11-1994
WO 0159151	A	16-08-2001	AU WO	3352101 A 0159151 A2		20-08-2001 16-08-2001
WO 02053728	A	11-07-2002	WO	02053728 A2		11-07-2002

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CA 02/00087

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. Claims Nos.: because they relate to subject matter not required to be searched by this Authority, namely:

2. Claims Nos.: because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210

3. Claims Nos.: because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this International application, as follows:

1. As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- The additional search fees were accompanied by the applicant's protest.
- No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210**Continuation of Box I.2**

Claims Nos.: 1,2,5-17,19-34, 37-58,70-81 (completely),
3,4,18,35,36,59-69,82-106 (partially)

Claims 1,2,5-17,19-34,37-58,70-81 do not satisfy the requirement of clarity (Article 6 PCT), because it is an undue burden for the skilled person to determine whether or not a given composition falls within the scope of said claims. Moreover, the claims define the subject-matter for which protection is sought by the result to be achieved and merely set out the problem to be solved without defining the technical features essential for the solution of the problem. This argument applies to claims 6-16,26-31,37-41 which require certain degrees of hybridization under a defined set of conditions. The composition cannot be defined by method features and no set of conditions is defined in any of said claims and it is an undue burden for the skilled person to determine the set of conditions.

The claims lack thus clarity to an extent that no meaningful search for claims 1,2,5-17,19-34,37-58,70-81 is possible.

Claim 3 relates to a composition comprising a set of oligonucleotides that is based on the sequences tested in Example 2 as set out in Table IA. In Example 2 100 oligonucleotides were tested (application page 53 Table I third column). Consequently, an extremely large number of possible compositions are covered by claim 3 such that a lack of clarity within the meaning of Article 6 PCT arises to such an extent as to render a meaningful search of the claim impossible.

Therefore, the search for claim 3 has been restricted to a composition comprising oligonucleotides selected from the set of oligonucleotides 1-10 assigned in Table I third column (page 53ff) that were tested in Example 2 (i.e. SEQ ID NOS 1-7,9,13,174). The search for claims 3,4,18,35,36,59-69,82-106 has been restricted, accordingly.

The applicant's attention is drawn to the fact that claims, or parts of claims, relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/CA 02/00087

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, X	WO 01 59151 A (TM BIOSCIENCE CORP ; JANOTA VIT (CZ); BULLOCK RICHARD S (US); PANCO) 16 August 2001 (2001-08-16) cited in the application the whole document	3, 4, 18, 35, 36, 59-69, 82-91
E	WO 02 053728 A (BOONE CHARLES ; BUSSEY HOWARD (CA); JIANG BO (CA); ROEMER TERRY (CA) 11 July 2002 (2002-07-11) SEQ ID NOS: 2424, 3783, 2892, 3126, 1477	3, 4, 18, 35, 36, 59-69, 82-91

INTERNATIONAL SEARCH REPORT

International Application No
PCT/CA 02/00087

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 799 897 A (AFFYMETRIX INC) 8 October 1997 (1997-10-08) abstract page 2, line 36 -page 4, line 1 page 4, line 54 -page 5, line 14 page 8, line 56 -page 14, line 21; claims 1-25; figure 4; examples 2-4 —	3, 4, 18, 35, 36, 59-69, 82-91
X	US 5 654 413 A (BRENNER SYDNEY) 5 August 1997 (1997-08-05) cited in the application abstract column 4, line 9 - line 51 column 6, line 30 -column 9, line 21 column 27-30: Appendix I column 16, line 59 -column 19, line 54; table II —	3, 4, 18, 35, 36, 59-69, 82-91
X	EP 0 698 792 A (NISSUI SEIYAKU CO) 28 February 1996 (1996-02-28) abstract page 3, line 51 -page 6, line 33; figures 1-19; examples 1,2 —	3, 4, 18, 35, 36, 59-69, 82-91, 100-106
X	NIEMEYER CHRISTOF M ET AL: "DNA-directed immobilization: Efficient, reversible, and site-selective surface binding of proteins by means of covalent DNA-streptavidin conjugates" ANALYTICAL BIOCHEMISTRY, ACADEMIC PRESS, SAN DIEGO, CA, US, vol. 268, no. 1, 1 March 1999 (1999-03-01), pages 54-63, XP002176566 ISSN: 0003-2697 abstract page 60, column 2, paragraph 2 -page 62, column 2, paragraph 1; figure 1 —	3, 4, 18, 35, 36, 59-69, 82-91, 100-106
A	SOUTHERN E ET AL: "MOLECULAR INTERACTIONS ON MICROARRAYS" NATURE GENETICS, NEW YORK, NY, US, vol. 21, no. SUPPL, January 1999 (1999-01), pages 5-9, XP000865979 ISSN: 1061-4036 cited in the application page 7, column 2, paragraph 2 -page 8, column 2, paragraph 2 —	3, 4, 18, 35, 36, 59-69, 82-106

-/-

INTERNATIONAL SEARCH REPORT

International Application No
PCT/CA 02/00087

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the International search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, PAJ, EMBASE, MEDLINE, SEQUENCE SEARCH

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	<p>WO 00 58516 A (WHITEHEAD BIOMEDICAL INST ; AFFYMETRIX INC (US)) 5 October 2000 (2000-10-05)</p> <p>abstract page 2, line 1 -page 3, line 22; figures 2-7 page 24, line 1 -page 26, line 23; claims 1-3,21-25; examples 5-7; table 1</p> <p style="text-align: center;">-/-</p>	<p>3, 4, 18, 35, 36, 59-69, 82-103, 105, 106</p>

Further documents are listed in the continuation of box C.

Patent family members are listed in annex.

* Special categories of cited documents :

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

- *T* later document published after the International filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- *&* document member of the same patent family

Date of the actual completion of the international search

16 April 2003

Date of mailing of the international search report

25/04/2003

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Tilkorn, A-C

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

**(19) World Intellectual Property Organization
International Bureau**



A standard linear barcode is located at the bottom of the page, spanning most of the width.

**(43) International Publication Date
1 August 2002 (01.08.2002)**

PCT

**(10) International Publication Number
WO 02/059354 A3**

(51) International Patent Classification:	C12Q 1/68	(74) Agents: HUNT, John, C. et al.; Blake, Cassels & Graydon LLP, Box 25, Commerce Court West, Toronto, Ontario M5L 1A9 (CA).
(21) International Application Number:	PCT/CA02/00087	
(22) International Filing Date:	25 January 2002 (25.01.2002)	
(25) Filing Language:	English	(81) Designated States (<i>national</i>): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.
(26) Publication Language:	English	
(30) Priority Data:		(84) Designated States (<i>regional</i>): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
(71) Applicant (<i>for all designated States except US</i>):	TM BIO-SCIENCE CORPORATION [CA/CA]; 439 University Avenue, Suite 1100, Toronto, Ontario M5G 1Y8 (CA).	
(72) Inventors; and		
(75) Inventors/Applicants (<i>for US only</i>):	PANCOSKA, Petr [CZ/US]; 901 Hinman Avenue #2C, Evanston, IL 60202 (US). JANOTA, Vit [CZ/CZ]; Ovenceka 27, 170 00 Praha 7 (CZ). BENIGHT, Albert, S. [US/US]; 1630 Valley View Drive, Schaumburg, IL 60193 (US). BULLOCK, Richard, S. [US/US]; 3500 North Lake Shore Drive, Chicago, IL 60657 (US). RICCELLI, Peter, V. [US/US]; 16830 Richards Drive, Tinley Park, IL 60477 (US). KOBLER, Daniel [CH/CA]; 33 Wood Street, Apartment 1102, Toronto, Ontario M4Y 2P8 (CA). FIELDHOUSE, Daniel [CA/CA]; 7 Chaplin Court, Bolton, Ontario L7E 5Y1 (CA).	Published: — with international search report — before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments
(88) Date of publication of the international search report:		26 June 2002

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

- with international search report
 - before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

(88) Date of publication of the international search report:
26 June 2003

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/059354 A3

(54) Title: POLYNUCLEOTIDES FOR USE AS TAGS AND TAG COMPLEMENTS, MANUFACTURE AND USE THEREOF

(57) Abstract: A family of minimally cross-hybridizing nucleotide sequences, methods of use, etc. A specific family of 210 24mers is described.

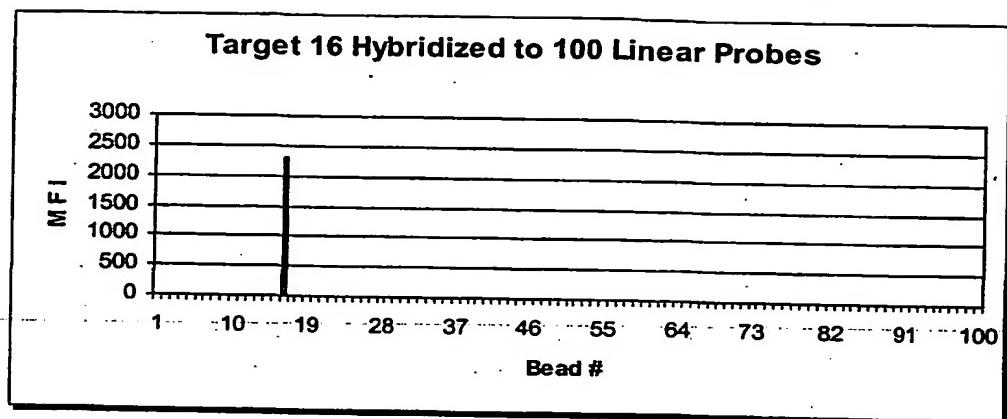


Figure 4

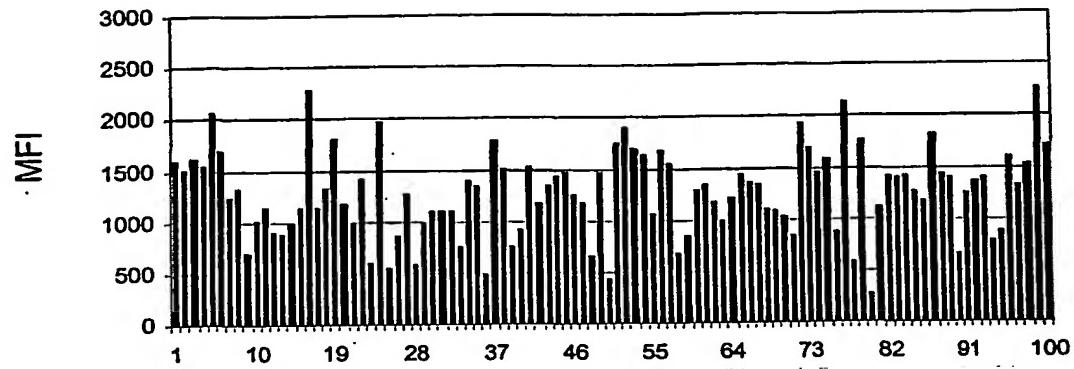


Figure 2

Best Available Copy

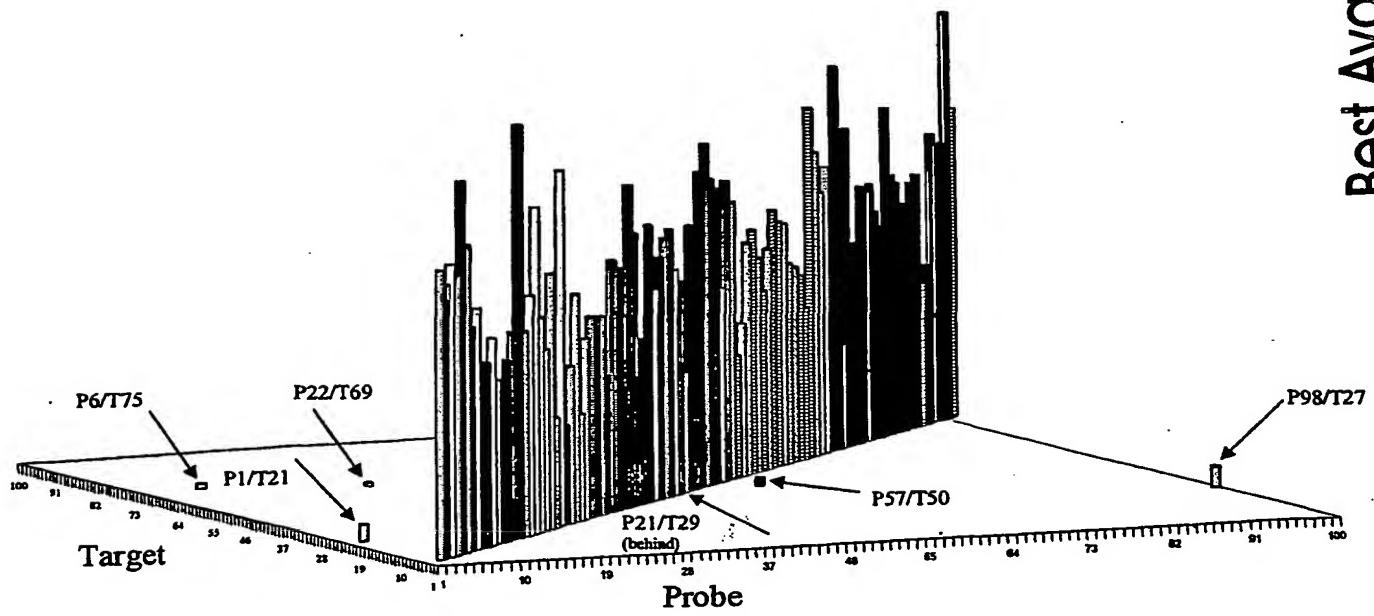


Figure 3